AAMAS-11 Tutorial

Decision Making in Multiagent Settings

Part I

Prashant Doshi and Zinovi Rabinovich

Part II

Matthijs Spaan and Christopher Amato

World of catastrophes

- Nature
  - 2004/12/26 – Sumatra-Andaman Earthquake
    - Magnitude estimate between 9.1 and 9.3
    - Triggered tsunamis causing 230,000 fatalities
  - 2005/08 – Hurricane Katrina
    - 1,836 dead
    - $81.2 billion damage
- Human
  - 26 April 1986 – Chernobyl atomic reactor meltdown
  - 11 September 2001 – Twin Towers in New York

Catastrophes: science

- Great Hanshin earthquake (1995). Killed over 6,400 people in and around Kobe, Japan.
- The data served to prototype a rescue simulation: **Robocup Rescue Domain**
  - Captures the dynamics of natural and man factor disasters and civil disorders
    - Includes uncertainty of various parameters
  - Realistically simulates the events: fire, traffic, building collapses, road blockage, etc.

Robocup Rescue - Scenario

- Given a post-event situation
  - Civilians trapped under collapsed buildings, and their life signs weakening with time
  - Some access routs are blocked or destroyed
  - Fires and civil disorder start and spread throughout the event site
- Manage platoons of Fire brigades, Police forces and Ambulance teams
  - Save as many people as possible
  - Recover and preserve site and its infrastructure (buildings, communications, etc.)

Robocup Rescue - Elements

- General capabilities
  - Mobility, communication, partial situation awareness at higher reasoning levels
- Specialisations
  - Ambulance teams rescue civilians from rubble and transport to safety
  - Fire brigades extinguishing fires
  - Police forces for traffic ordering, general order and safety
- Our Target: Provide automated decision and information support for *time critical* and potentially *irreversible* decisions.

Task 1: ambulance allocation

- Multiple ambulance services
  - Business oriented operation
  - Competition for government funds and public opinion
- Given several locations that require medical assistance, how many ambulances from which firm will go to which location?

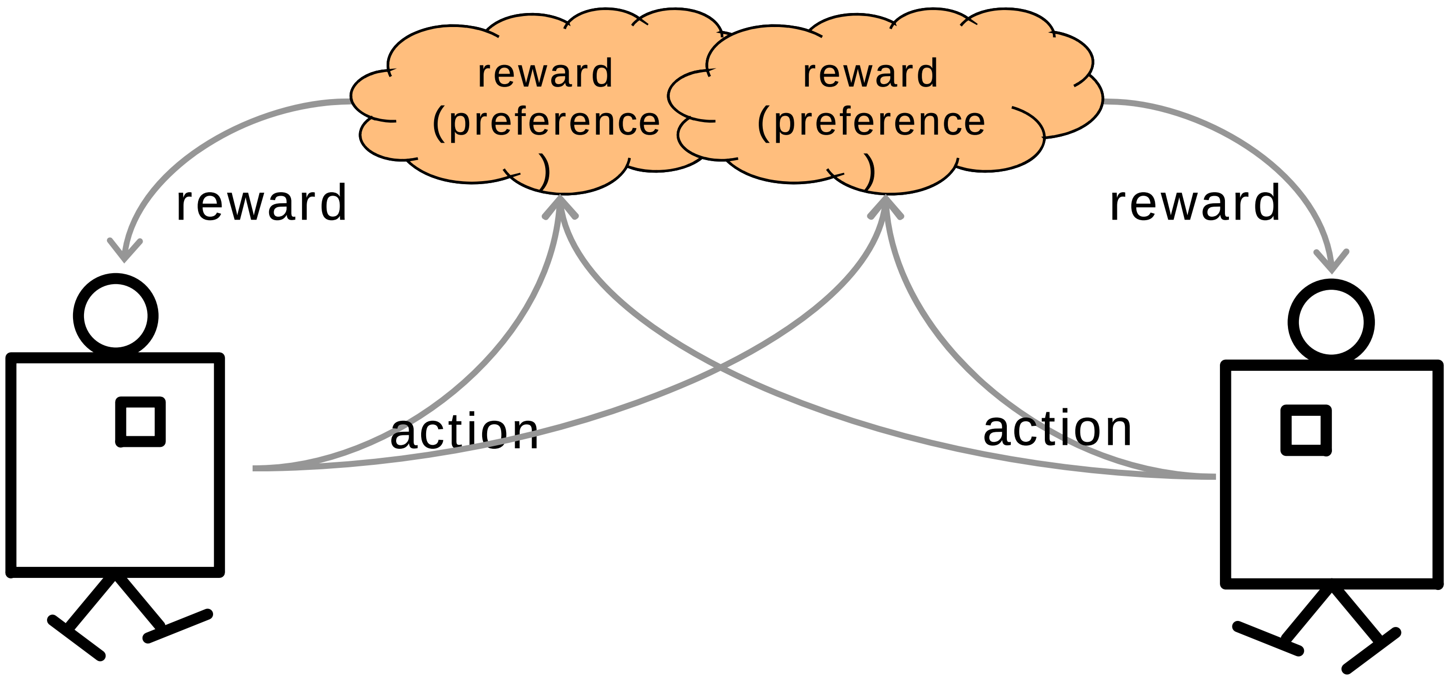
Task 2: police patrols

- Low ratio of police force vs. operative requirements
- How frequently and with what qualitative force to patrol an area?
- How many safe routs vs their quality can the given police force support? Can and should it be adapted over time?

Task 3: firefighters

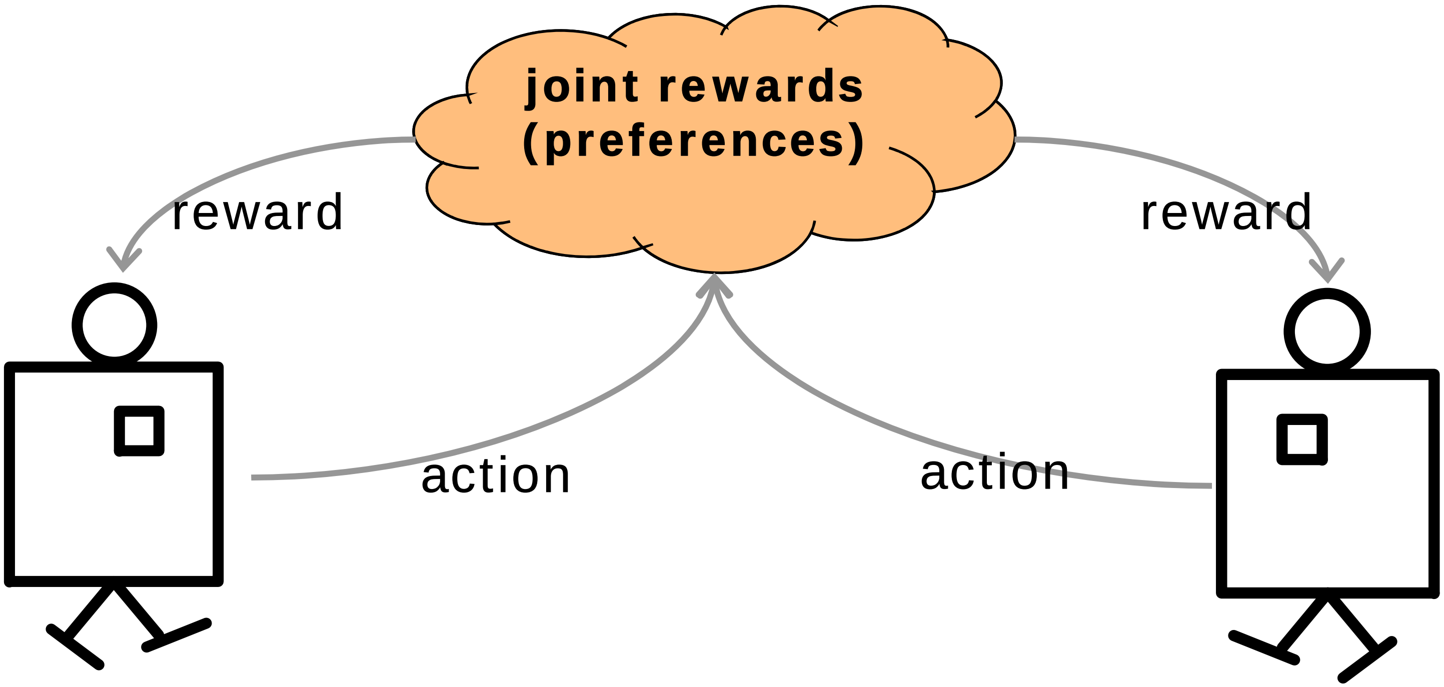
- Maintain effort toward saving the building or draw back and minimise the spread of fire?
- Concentrate on a multitude of smaller fires or allow controlled unification and deal with only one location?
  - Will transportation routs be endangered?
  - Are there still civilians evacuating from the area/building?
- Push through the fire to victims or save the fire crew and pull out?
  - If multiple crews are on site, which one goes? When?

Multiagent decision problem



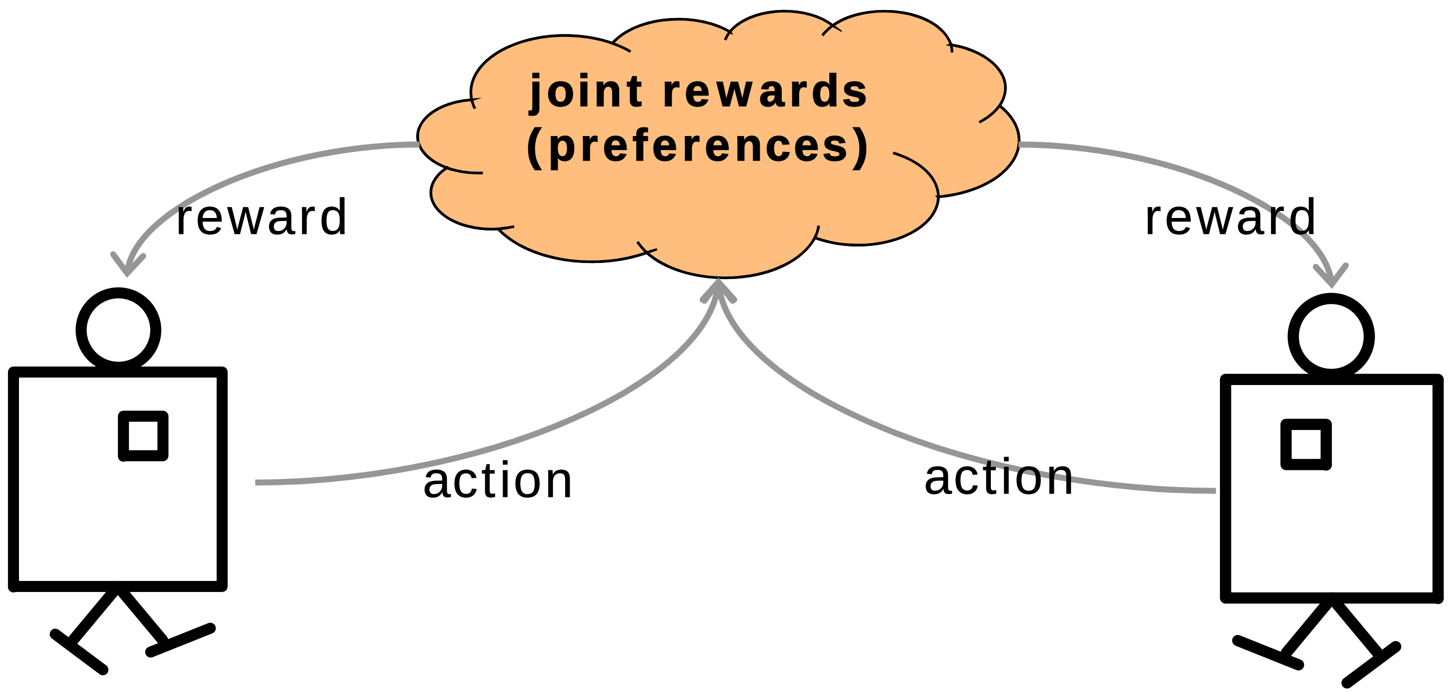
Each agent optimizes its rewards

Multiagent decision problem



Each agent optimizes its rewards

Multiagent decision problem

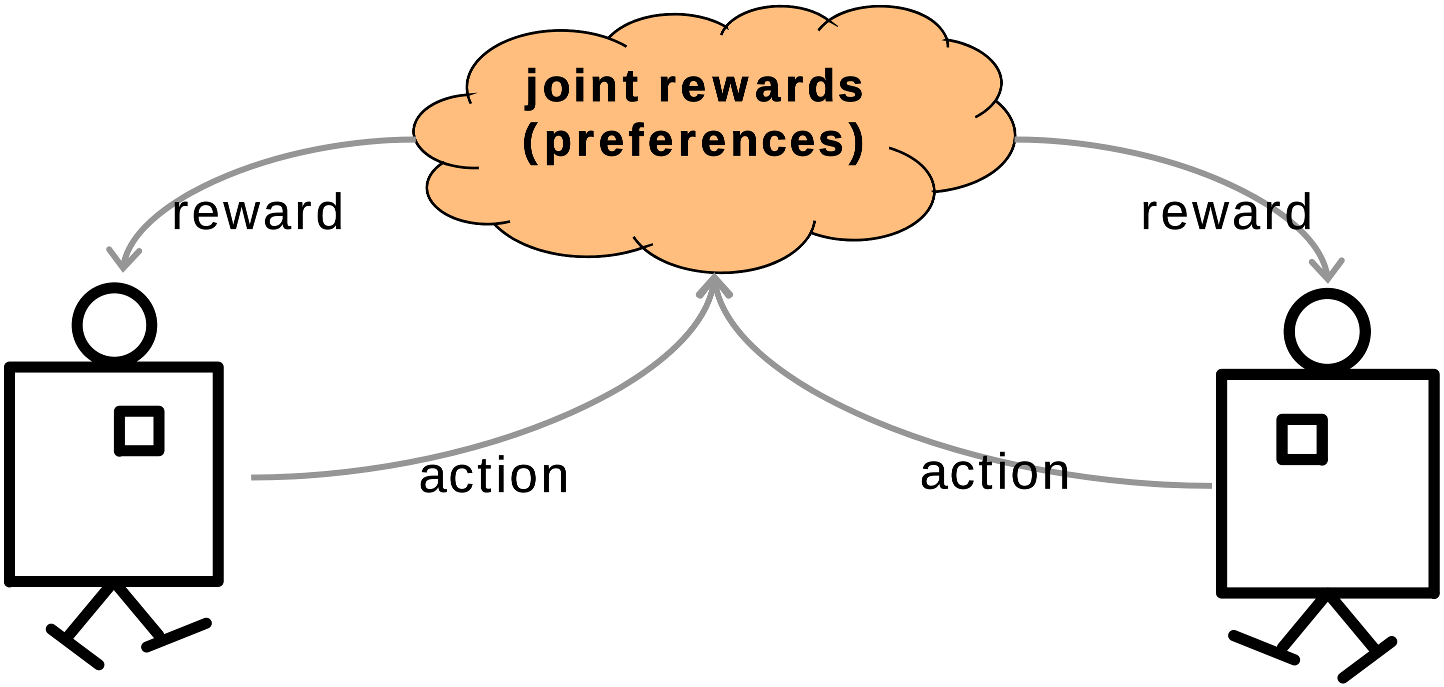


Each agent optimizes rewards

Single interaction *(game)*

Strategy: $\Delta(A)$

Multiagent decision problem

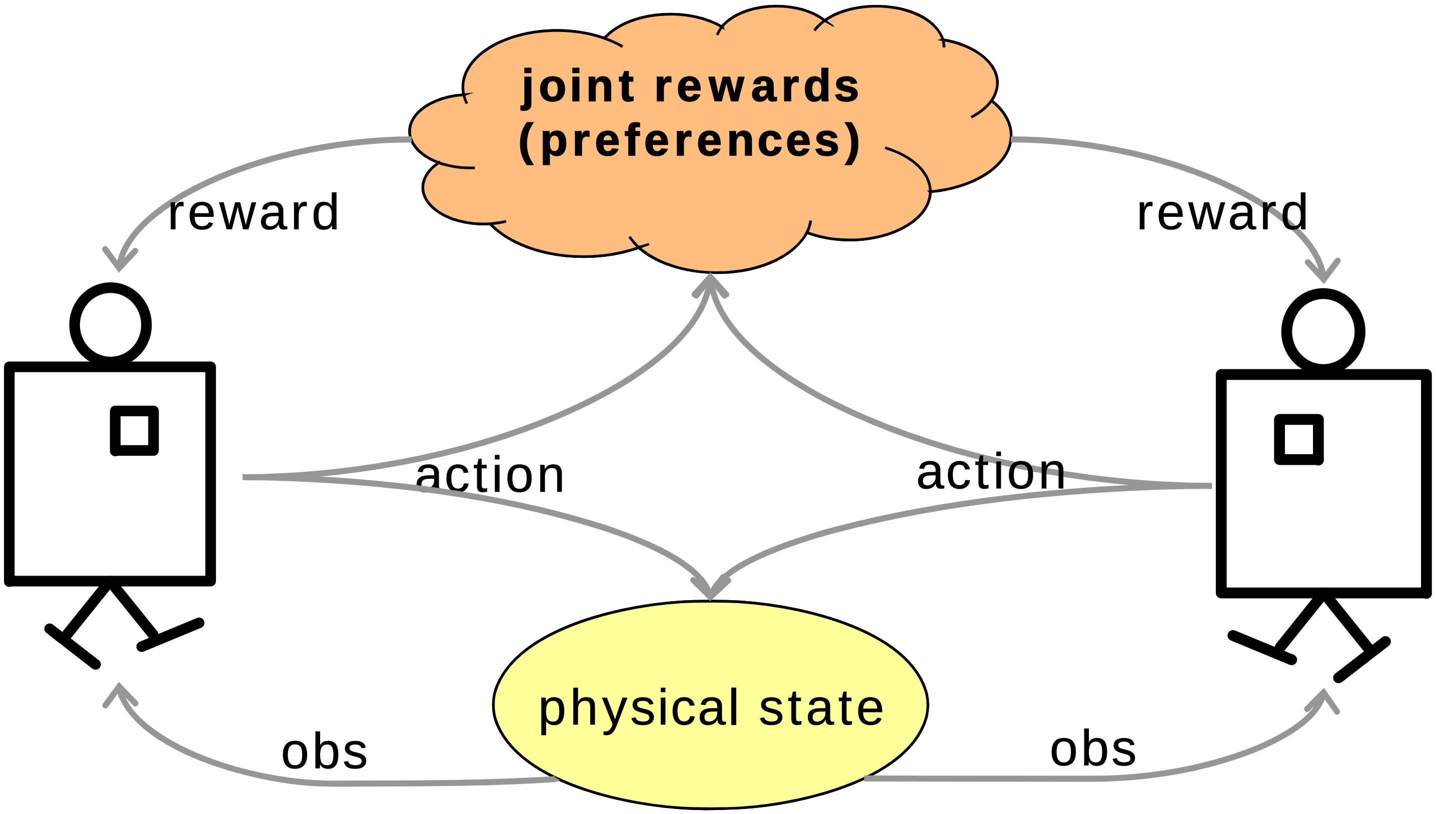


Each agent optimizes rewards

Repeated interactions

Strategy: History of observations $\rightarrow \Delta(A)$

Multiagent decision problem

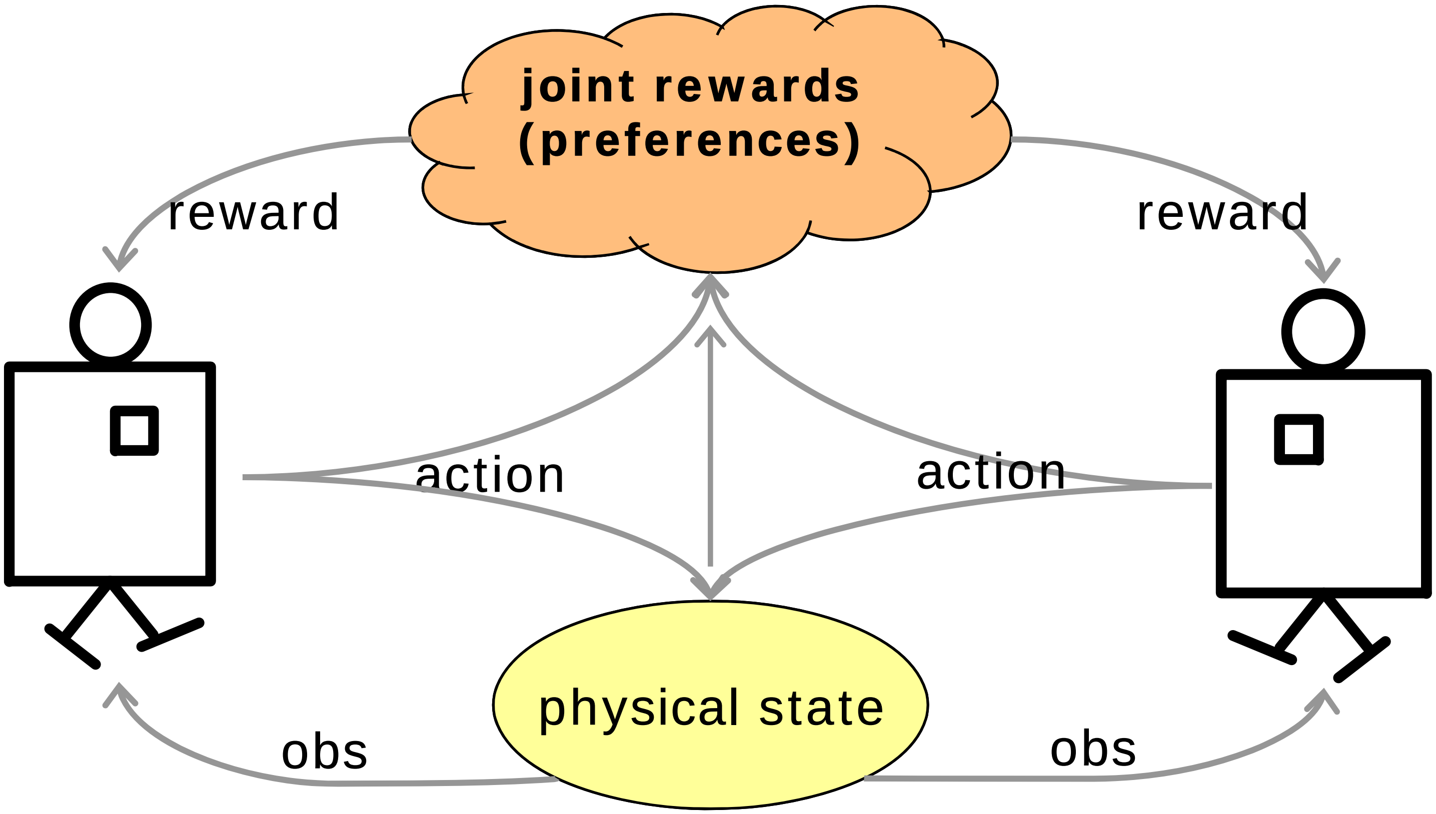


Each agent optimizes rewards

Repeated interactions

Strategy: History of observations $\rightarrow \Delta(A)$

Multiagent decision problem



Each agent optimizes rewards

Repeated interactions

Strategy: History of observations $\rightarrow \Delta(A)$

Dimensions of interaction

- Single or Extended
  - Strategies in extended interactions may be different
    - Extended: Finite or infinite interactions

- Cooperative or Non-cooperative

Dimensions of interaction

- Joint reward or Joint reward and state
  - State is dynamic, influenced by actions
  - State may influence rewards as well
- Perfect or Incomplete information about others

Predictive and epistemological requirements of solution

- In order to maximize rewards, predict actions of others
  - Common knowledge of rationality
    - All agents are rational; All know that all are rational; All know that all know that all are rational; ...
  - Common and perfect knowledge of rewards
    - All know others' rewards; All know that all know others' rewards; ...
  - Common and partial knowledge of rewards
    - Probability distribution over possible rewards is common knowledge

Predictive and epistemological requirements of solution

Epistemological requirements for rational behavior are strict!

Models of interactions (first glance)

Single and repeated interactions with joint rewards are the focus of traditional game theory

Interactions involving joint state and reward are the focus of decision theory inspired approaches to game theory. These generally include extensions of single agent decision-theoretic models to multiagent settings

Other applications

- Robotics
  - Planetary exploration

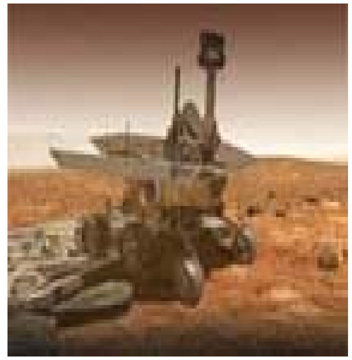
    Surface mapping by rovers

    Coordinate to explore pre-defined region optimally

    Uncertainty due to sensors
  - Robot soccer

    Coordinate with teammates and deceive opponents

    Anticipate and track others' actions



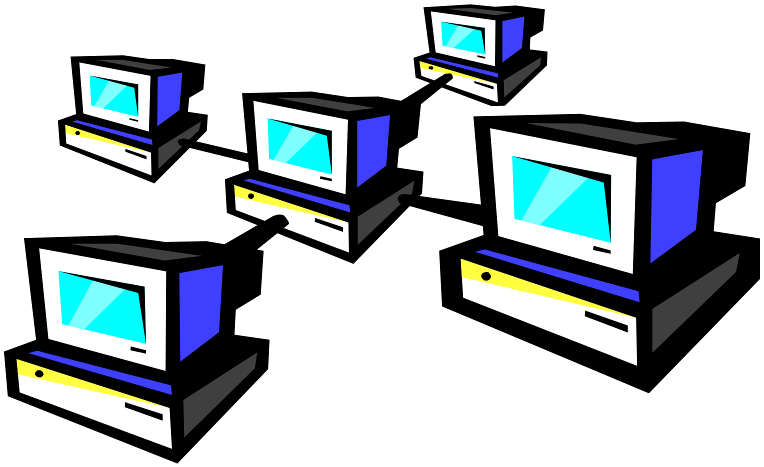
Spirit

Opportunity

RoboCup Competition

Other applications

- Defense
  - Coordinate UAV movements in battlefields
    - Exact “ground situation” unknown
  - Coordinate anti-air defense units



- Distributed Systems
  - Networked Systems
    - Packet routing
  - Sensor networks

Classroom game: Prisoner's dilemma

- Instructions

We are going to play a card game in which everybody will be matched with someone in the room. I will now give each of you a pair of playing cards, one red card (♥ or ♦) and one black card (♠ or ♣). The numbers or faces on the cards will not matter, just the color. You will be asked to play one of these cards by holding it to your chest . Your earnings are determined by the card that you play and by the card played by the person matched with you.

If you play your red card, then your earnings will increase by $2, and the earnings of the person matched with you will not change. If you play your black card, your earnings do not change and the earnings of the person matched with you go up by $3. If you each play your red card, you will each earn $2. If you each play the black card, you will each earn $3. If you play your black card and the other person plays his or her red card, then you earn zero and the other person earns the $5. If you play red and the other person plays black, you earn the $5, and the other person earns zero. All earnings are hypothetical. After you choose which card to play, hold it to your chest. We then tell you who you are matched with, and you can each reveal the card that you played. Record your earnings in the space below. To make this easier, please write your name: ____________________ .

To begin: Would the people in the row that I designate please choose which card to play and write the color (R or B) in the first column. Show that you have made your decision by picking up the card you want to play and holding it to your chest. Everyone finished? Now, I will pair you with another person, ask you to reveal your choice, and calculate your earnings. Remember to keep track of earnings in the space provided below. Finally, please note that in period 2 you will be matched with a different person, and payoffs will change. In period 3 you will be matched with a different person and payoffs change again, but you get to play with him/her in the last three periods.

Classroom game: Prisoner's dilemma

Your payoff table

Period	Your card (R or B)	Other's card (R or B)	Your earnings
1			
2			
3			
4			
5			

Classroom game: Prisoner's dilemma

Payoff table for Period 1

Player II

Player I

	black	red
black	3,3	0,5
red	5,0	2,2

Payoff table for Period 2

Player II

Player I

	black	red
black	8,8	0,10
red	10,0	2,2

Game in Normal Form

- Defined by a tuple $< I, \{A_i\}_{i \in I}, \{R_i\}_{i \in I} >$
  - I is the set of players, usually $I = \{1, ..., n\}$
  - A_i is the set of actions (*pure strategies*) available to player i.
    - Space of pure strategy profiles $A = \bigotimes_{i \in I} A_i$
    - Let $a = (a_i, a_{-i}) \in A$. Where $a_i \in A_i$ is the action prescribed to agent i, and $a_{-i} \in \bigotimes_{j \in I \setminus \{i\}} A_j = A_{-i}$ portion of profile adopted by other agents.
  - $R_i : A \to \mathcal{R}$ is the reward (*utility*) of the player i, given that players *simultaneously* play their actions
- Each agent *rationally* seeks to maximise its utility

Roadmap: Why game is a game?

- Is there a **guarantee** of utility if I don’t know how others act?
- If I know how others act, how should I?
  - What if I can guess, but not certain?
- If the game is to be repeated, should I act differently?

Guarantees

- "Enemy assumption": A player assumes that all others collude against it.
  - Essentially a zero sum game
    - $I = 1, 2$, and $R_1 = -R_2$.
  - Guarantee is $\max_{a_1 \in A_1} \min_{a_2 \in A_2} R_1(a_1, a_2)$
- Simplest example: Fire station location

Guarantees: example

- Two plants A and B build a new private fire station
  - Where should it be located?
- Assume fires are deliberate, then time of arrival dictates utility for the Fire Brigade:

		Fire at		
		A	A and B	B
Station	near A	0	-1	-1
	middle	-0.5	-0.5	-0.5
	near B	-1	-1	0

- Minimax value is -0.5 and minimax strategy is *middle*

Equilibria

- Given a partial profile $a_{-i} \in A_{-i}$ the action choice of agents except $i \in I$.
- a_i^* is a best response of agent $i \in I$ to a_{-i} if
  $a_i^* \in \arg\max_{a_i \in A_i} R_i(a_i, a_{-i})$
- A *strategy profile* (joint action) $a \in A$ is a pure Nash equilibria if for all $i \in I$ a_i is a best response to a_{-i}.

Equilibria: example

- Two plants A and B build a new private fire station. Where should it be located?
- Assume fires are deliberate, then time of arrival dictates utility for the Fire Brigade:

	A	A and B	B
near A	0	-1	-1
middle	-0.5	-0.5	-0.5
near B	-1	-1	0

- The pair (*A and B*, *middle*) is a pure Nash equilibria

Non-existence of pure Nash

- Police sends patrols to plant A and plant B to try and catch the saboteurs.
- Utility is determined by the similarity of actions:

	A	B
A	1	-1
B	-1	1

- It is easy to see that no pair $(a_{police}, a_{saboteur})$ is an equilibrium profile.
- Intuition: Surprise factor by randomisation

Mixed profile

- *Mixed strategy* of an agent $i \in I$ is a probability distribution π_i over A_i, where $\pi(a_i)$ is the probability of selecting action a_i.
- Denote Δ_i the set of all probability distributions over A_i. *Mixed strategy profile* (joint mixed strategy) is a distribution $\pi = (\pi_i, \pi_{-i}) \in \bigotimes_{i \in I} \Delta_i$.
  - $\pi(a) = \prod_{i \in I} \pi_i(a_i)$ is the probability that agents will jointly select pure profile $a \in A$.
  - *Expected utility* is then $E_\pi[R_i] = \sum_{a \in A} \pi(a) R_i(a)$

Mixed Nash equilibrium

- Given partial mixed profile π_{-i}. π_i^* is a best response mixed strategy if $\pi_i^* \in \arg\max_{\pi_i \in \Delta_i} E_{(\pi_i, \pi_{-i})}[R_i]$
- A complete mixed profile π is in *mixed Nash equilibrium* if for all $i \in I$, π_i is a best response to π_{-i}.
- For the police patrol example equally probable choice is an equilibrium.

Sad example

- Ambulances are independent business services
  - Cost driven and competitive
- Government funds:
  - Distributed in proportion to saved lives
  - Recognition for success in major events
- Scenario:
  - Two ambulance services
  - Three events: two are minor one major
    - Minor events are local to the services
    - Major event necessitates both services to handle

Sad example (cont)

- Assume that total government funds are 4 units
- If the major event is handled extra 2 units are allocated
- The utilities can be summarised by:

	Major	Minor
Major	(3,3)	(0,4)
Minor	(4,0)	(2,2)

- Problem: It is always best to handle the minor event.
- But in real life they do concentrate on major events. Why?

Repeated games

- Ambulance services "play" this game repeatedly.
  - Long term accumulation of utility
  - For infinite repetition discounting by $\gamma < 1$ or averaging of a single repetition utility, r_i^t, are used.

$$\sum_{t=1}^{\infty} \gamma^t r_i^t \text{ or } \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} r_i^t$$

- Sequences of actions (or rules composing them) are considered
  - Behaviour rules producing action sequences are termed *policy*
  - In presence of memory new possibilities occur: trust, revenge, reciprocity, etc.

Happy example

- Consider again:

	Major	Minor
Major	(3,3)	(0,4)
Minor	(4,0)	(2,2)

- Assume the following *tit-for-tat* policy:
  - At first attempt to choose “Major”
  - Then mimic the previous action of the other agent
- It is easy to see that TFT is an equilibrium for infinite utility accumulation, and that (Major, Major) is infinitely repeated.

Bayesian games

- Relax the assumption of perfect knowledge of agents’ rewards

- Type system
  - Agent’s type: Encompasses private information relevant to the agent’s behavior
  - Joint probability distribution over types, which is common knowledge

Bayesian games

In Harsanyi's own words:

". . . we can regard the attribute vector c_i *as representing certain physical, social, and psychological attributes of player i himself in that it summarizes some crucial parameters of player i's own payoff function* U_i *as well as main parameters of his beliefs about his social and physical environment . . ."*

Bayesian games – Example

Criminals

Police Patrol

	Enter	Stay out
Enter	0,-1	2,0
Stay out	2,1	3,0

Policing is weak

Criminals

	Enter	Stay out
Enter	1.5,-1	3.5,0
Stay out	2,1	3,0

Policing is strong

Type space: $\Theta_{Police} = \{R_{Weak}, R_{Strong}\}$

Bayesian games – Example

Policing is weak

Police Patrol \ Criminals	Enter	Stay out
Enter	0,-1	2,0
Stay out	2,1	3,0

Policing is strong

Police Patrol \ Criminals	Enter	Stay out
Enter	1.5,-1	3.5,0
Stay out	2,1	3,0

Let p be the probability that the police is weak

	Enter	Stay out
Enter, Enter	1.5(1-p),-1	2p+3.5(1-p),0
Enter, Stay out	2(1-p), -p+(1-p)	2p + 3(1-p),0
Stay out, Enter	2p + 1.5(1-p), p – (1-p)	3p + 3.5(1-p),0
Stay out, Stay out	2,1	3,0

Bayesian games – Example

Criminals

Police Patrol	**Enter**	**Stay out**
Enter	0,-1	2,0
Stay out	2,1	3,0

Policing is weak

Criminals

	Enter	**Stay out**
Enter	1.5,-1	3.5,0
Stay out	2,1	3,0

Policing is strong

For all $p \geq 0$, (Enter, Enter) and (Enter, Stay out) is dominated

	Enter	**Stay out**
Enter, Enter	1.5(1-p),-1	2p+3.5(1-p),0
Enter, Stay out	2(1-p), -p+(1-p)	2p + 3(1-p),0
Stay out, Enter	2p + 1.5(1-p), p – (1-p)	3p + 3.5(1-p),0
Stay out, Stay out	2,1	3,0

Bayesian games – Example

Criminals (Police Patrol rows)

	Enter	**Stay out**
Enter	0,-1	2,0
Stay out	2,1	3,0

Policing is weak

Criminals (Police Patrol rows)

	Enter	**Stay out**
Enter	1.5,-1	3.5,0
Stay out	2,1	3,0

Policing is strong

For all p ≥ 0, (Enter, Enter) and (Enter, Stay out) is dominated

so the games collapses into:

	Enter	**Stay out**
Stay out, Enter	2p + 1.5(1-p), p – (1-p)	3p + 3.5(1-p),0
Stay out, Stay out	2,1	3,0

Bayesian games – Example

Criminals (Policing is weak)

Police Patrol	Enter	Stay out
Enter	0,-1	2,0
Stay out	2,1	3,0

Policing is weak

Criminals (Policing is strong)

Police Patrol	Enter	Stay out
Enter	1.5,-1	3.5,0
Stay out	2,1	3,0

Policing is strong

	Enter	Stay out
Stay out, Enter	1.5 + 0.5p, 2p -1	3.5 – 0.5p, 0
Stay out, Stay out	2,1	3,0

For $p > 0.5$, Enter is a dominating action for the criminal and {(Stay out, Stay out),Enter} is a Nash equilibrium

For $p \leq 0.5$, {(Stay out, Stay out), Enter} and {(Stay out, Enter), Stay out} are Nash equilibria

Bayesian games – Example

Criminals

Police Patrol		Enter	Stay out
	Enter	0,-1	2,0
	Stay out	2,1	3,0

Policing is weak

Criminals

	Enter	Stay out
Enter	1.5,-1	3.5,0
Stay out	2,1	3,0

Policing is strong

	Enter	Stay out
Stay out, Enter	1.5 + 0.5p, 2p -1	3.5 – 0.5p, 0
Stay out, Stay out	2,1	3,0

EU(Stay out, Enter) = $(1.5+0.5p)x+(1-x)(3.5-0.5p)=3.5-0.5p+x(p-2)$

EU(Stay out, Stay out) = $2x+3(1-x)=3-x$

Police is indifferent when $3.5p-0.5p+x(p-2)=3-x$

$x=1/2$

Bayesian games – Example

Policing is weak

Police Patrol \ Criminals	Enter	Stay out
Enter	0,-1	2,0
Stay out	2,1	3,0

Policing is strong

Police Patrol \ Criminals	Enter	Stay out
Enter	1.5,-1	3.5,0
Stay out	2,1	3,0

	Enter	**Stay out**
Stay out, Enter	1.5 + 0.5p, 2p -1	3.5 – 0.5p, 0
Stay out, Stay out	2,1	3,0

EU(Enter) = $(2p-1)y+1(1-y)=(2p-2)y+1$

EU(Stay out) = 0

Criminal is indifferent when $1+y(2p-2)=0$

$$y=1/2(1-p)$$

Bayesian games – Example

Policing is weak

Police Patrol \ Criminals	Enter	Stay out
Enter	0,-1	2,0
Stay out	2,1	3,0

Policing is strong

Police Patrol \ Criminals	Enter	Stay out
Enter	1.5,-1	3.5,0
Stay out	2,1	3,0

3 Bayesian Nash equilibria

{Stay out, Enter} for any p

{(Stay out, Enter), Stay out} if $p \leq 0.5$

$$\left\{\left\langle \frac{1}{2(1-p)}, \frac{1-2p}{2(1-p)} \right\rangle, \left\langle \frac{1}{2}, \frac{1}{2} \right\rangle\right\}$$ if $p \leq 0.5$

Bayesian games

In general, a strategy profile $\{\pi_i, \pi_j\}$ is a Bayesian Nash equilibrium if for each agent i and its type, θ_i ,

$$\pi_i(\theta_i) = \operatorname*{argmax}_{a_i \in A_i} \sum_{\theta_j \in \Theta_j} R_{\theta_i}(a_i, \pi_j(\theta_j)) p(\theta_i, \theta_j)$$

Repeated games

In game theory, two models of decision-making in repeated interactions are popular:

- Fictitious play
- Rational learning

Repeated games – Fictitious play

- Simplest model of decision-making in repeated games
- At each stage, an agent ascribes a mixed strategy to the other, $b_i^t(a_j)$

  Other agent is assumed to act according to this mixed strategy
- The strategy is computed as follows:

$$F^t(a_j) = F^{t-1}(a_j) + \begin{cases} 1 & if \;\; a_j^{t-1} = a_j \\ 0 & if \;\; a_j^{t-1} \neq a_j \end{cases}$$

$$b_i^t(a_j) = \frac{F^t(a_j)}{\sum_{a_j \in A_j} F^t(a_j)}$$

Maintain a frequency count of previous actions

- Agent computes its best response to the mixed strategy of other

Fictitious play – Example

Police patrol 2

		Enter	Stay out
Police patrol 1	Enter	0,0	1,1
	Stay out	1,1	0,0

coordination game

2 pure strategy Nash equilibria and one mixed strategy Nash equilibrium

{Enter, Stay out} {Stay out, Enter}

$$\left\{\left\langle 0.5, 0.5\right\rangle, \left\langle 0.5, 0.5\right\rangle\right\}$$

Fictitious play – Example

Police patrol 2

Police patrol 1	Enter	Stay out
Enter	0,0	1,1
Stay out	1,1	0,0

coordination game

Round	Patrol 1	Patrol 2	1's belief	2's belief
0			(1,0.5)	(1,0.5)
1	Stay out	Stay out	(1,1.5)	(1,1.5)
2	Enter	Enter	(2,1.5)	(2,1.5)
3	Stay out	Stay out	(2,2.5)	(2,2.5)
4	Enter	Enter	(3,2.5)	(3,2.5)
…	…	…	…	…

Fictitious play – Example

Police patrol 2

Police patrol 1	Enter	Stay out
Enter	0,0	1,1
Stay out	1,1	0,0

coordination game

Round	Patrol 1	Patrol 2	1's belief	2's belief
0			(1,0.5)	(1,0.5)
1	Stay out	Stay out	(1,1.5)	(1,1.5)
2	Enter	Enter	(2,1.5)	(2,1.5)
3	Stay out	Stay out	(2,2.5)	(2,2.5)
4	Enter	Enter	(3,2.5)	(3,2.5)
…	…	…	…	…

Nash equilibrium!

Fictitious play

Interesting properties

- If an action vector is a strict Nash equilibrium of a stage game, it is the steady state of fictitious play in the repeated game
- If the empirical distribution of each agent's strategies converges in fictitious play, then it converges to a Nash equilibrium
- Fictitious play in repeated games converges if the game is a 2x2 game with generic payoffs or is a zero-sum game

Roadmap: Stochastic Games

- Games become increasingly general
  - Some interaction parameters can be uncertain
    - E.g. in Bayesian Equilibria the reward
  - Interaction can be extended over time
    - E.g. in FP a long term reward average was used
- What other properties can be generalised?
  - A repeated game can have a state
    - E.g. the amount of water firefighters have
  - A game can be partially observed (monitored)
    - E.g. fumes and smoke conceal the actual fire

Markovian Environment

- Consider the tuple $< S, s_0, A, T >$
  - S set of agent's world states, with s_0 being the initial one
  - A is the set of actions available to the agent
  - $T : S \times A \times S \longrightarrow [0, 1]$ is the transition matrix. $T(s', a, s)$ is the probability that the world will change from state $s \in S$ to state $s' \in S$ if agent performs $a \in A$
- What a rational agent would do with such a setting?

How does it work?

- At time $t = 0$ the world starts at state s_0
- Then decision loop is repeated
  - Agent chooses an action $a_t \in A$
  - Action a_t is applied
  - The world changes its state. s_{t+1} is chosen w.r.t. $T(\cdot|s_t, a_t)$
  - Time step occurs $t \leftarrow t + 1$
- How does an agent choose its action?

Example

- For example the crime rate is weakly responsive to the police presence
- Modelled by a Markovian environment
  - $S = \{high, medium, low\}$ is the crime rate
  - $A = \{large, small\}$ is the police force size

$T(\cdot, a, \cdot)$	$a = large$			$a = small$		
	high	medium	low	high	medium	low
high	0	0.7	0.3	1	0	0
medium	0	0.5	0.5	0.5	0.5	0
low	0	0	1	0.1	0.3	0.6

Markov Decision Problem

- The tuple $< S, s_0, A, T >$ is only the *environment*
- Rational agents needs a performance measure to decide on an action (sequence)
- Markov Decision Problem (MDP) is a tuple $< S, s_0, A, T, r >$
  - Given a utility function $r : S \times A \times S \longrightarrow \mathbf{R}$
  - Utility based performance measure
    - Finite horizon $T < \infty$: $\mathbf{E}\left(\sum_{t=0}^{T} r(s_{t+1}, a_t, s_t)\right)$
    - Infinite horizon $\gamma < 1$: $\mathbf{E}\left(\sum_{t=0}^{\infty} \gamma^t r(s_{t+1}, a_t, s_t)\right)$
    - Infinite Average: $\lim_{T \to \infty} \mathbf{E}\left(\frac{1}{T} \sum_{t=0}^{T} r(s_{t+1}, a_t, s_t)\right)$

Action sequence by policy

- Formally infinite performance measures would require strategies to be infinite sequences of actions
- Instead we define a *policy*
  - Repeatedly applied rule to construct the sequence
  - We'll focus on $\pi : S \rightarrow \Delta(A)$, where $\Delta(A)$ is the space of distributions over A
- Sufficiency of policy space
  - The sufficient statistics set for previous activity is the domain
  - Performance may not be improved by a more complex policy
  - $\pi : S \rightarrow \Delta(A)$ is sufficient for single agent MDPs

How good is a policy?

- Denote $V^\pi(s)$ the utility accumulated by an agent following policy π if the system starts in state s.
$$V^\pi(s) = \sum_a \pi(s,a) \sum_{s'} \left(R(s',a,s) + \gamma V^\pi(s')\right) T(s'|s,a)$$
- Define auxiliary quality of action $Q^\pi(s,a)$
  - Denotes the utility gained by an agent by applying $a \in A$ in state s and then following policy π
$$V^\pi(s) = \sum_a \pi(s,a) Q^\pi(s,a)$$
$$Q^\pi(s,a) = \sum_{s'} \left(R(s',a,s) + \gamma V^\pi(s')\right) T(s'|s,a)$$
- Notice that given π, V^π is the solution to a system of linear equations

Example

- Crime rate model:
  - $S = \{high, medium, low\}$ is the crime rate
  - $A = \{large, small\}$ is the police force size

$T(\cdot, a, \cdot)$	$a = large$			$a = small$		
	high	medium	low	high	medium	low
high	0	0.7	0.3	1	0	0
medium	0	0.5	0.5	0.5	0.5	0
low	0	0	1	0.1	0.3	0.6

- Police chief will receive:
  - A reprimand if the crime rate increases
  - A frown from his neighbour if it remains the same
  - A medal if it drops
  - A bad reputation if he uses too much force

Example

- Crime rate model:
  - $S = \{high, medium, low\}$ is the crime rate
  - $A = \{large, small\}$ is the police force size

$T(\cdot, a, \cdot)$	$a = large$			$a = small$		
	high	medium	low	high	medium	low
high	0	0.7	0.3	1	0	0
medium	0	0.5	0.5	0.5	0.5	0
low	0	0	1	0.1	0.3	0.6

- Police chief utility is:

$R(\cdot, a, \cdot)$	$a = large$			$a = small$		
	high	medium	low	high	medium	low
high	-1.5	0	0	-0.5	1	1
medium	-2	-1.5	0	-1	-0.5	1
low	-2	-2	-1.5	-1	-1	-0.5

Example

- A policy $\pi : S \to \Delta(A)$ for the chief would be to decide how many people he send out every day with what probability depending on that day's situation.
- Assume that he always send out large force $\pi(s) = (1, 0)$
- Assume also that he likes to say "Tomorrow is another day" and assigns $\gamma = 0.5$
- What would be his benefit?

Example

	$T(\cdot, a = large, \cdot)$			$R(\cdot, a = large, \cdot)$		
	high	medium	low	high	medium	low
high	0	0.7	0.3	-1.5	0	0
medium	0	0.5	0.5	-2	-1.5	0
low	0	0	1	-2	-2	-1.5

$$
\begin{aligned}
V^\pi(s) &= \sum_{s'} (R(s', a, s) + \gamma V^\pi(s'))T(s'|s, a) \\
V^\pi(h) &= 0.0 * (..) + 0.7 * (0.0 + 0.5V^\pi(m)) + \ldots \\
& \quad 0.3 * (0.0 + 0.5V^\pi(l)) \\
V^\pi(m) &= 0.0 * (..) + 0.5 * (-1.5 + 0.5 * V^\pi(m)) + \ldots \\
& \quad 0.5 * (0.0 + 0.5V^\pi(l)) \\
V^\pi(l) &= 0.0 * (..) + 0.0 * (..) + 1.0 * (-1.5 + 0.5V^\pi(l))
\end{aligned}
$$

Example

	$T(\cdot, a = large, \cdot)$			$R(\cdot, a = large, \cdot)$		
	high	medium	low	high	medium	low
high	0	0.7	0.3	-1.5	0	0
medium	0	0.5	0.5	-2	-1.5	0
low	0	0	1	-2	-2	-1.5

$$
\begin{aligned}
V^\pi(s) &= \sum_{s'} (R(s', a, s) + \gamma V^\pi(s'))T(s'|s, a) \\
V^\pi(h) &= 0.35V^\pi(m) + 0.15V^\pi(l) \\
V^\pi(m) &= -0.75 + 0.25V^\pi(m) + 0.25V^\pi(l) \\
V^\pi(l) &= -1.5 + 0.5V^\pi(l)
\end{aligned}
$$

Example

	$T(\cdot, a = large, \cdot)$			$R(\cdot, a = large, \cdot)$		
	high	medium	low	high	medium	low
high	0	0.7	0.3	-1.5	0	0
medium	0	0.5	0.5	-2	-1.5	0
low	0	0	1	-2	-2	-1.5

$$\begin{aligned} V^\pi(h) &= -1.15 \;\; (\max \approx -0.59) \\ V^\pi(m) &= -2 \;\; (\max \approx -1.13646) \\ V^\pi(l) &= -3 \;\; (\max \approx -1.285714) \end{aligned}$$

Optimal policy

- Rational agent would like to find $\pi^* \in \arg\max_\pi V^\pi(s_0)$
- Bellman-Ford Equation:
  - Exists V^* so that:

$$V^*(s) = \max_\pi \sum_a \pi(s,a) \sum_{s'} \left(R(s',a,s) + \gamma V^*(s')\right) T(s'|s,a)$$

  - $V^* = \max_\pi V^\pi$, and exists π^* so that $V^* = V^{\pi^*}$

$$\pi^*(s,\cdot) = \arg\max_{\pi(s,\cdot)} \sum_a \pi(s,a) \sum_{s'} \left(R(s',a,s) + \gamma V^*(s')\right) T(s'|s,a)$$

- But how do we find V^*??

Value Iteration

- Dynamic Programming solution
  - Start from some arbitrary small $V_0(\cdot)$
  - Propagate back in time:

$$V_{t+1}(s) = \max_{\pi} \sum_{a} \pi(s,a) \sum_{s'} \left(R(s',a,s) + \gamma V_t(s')\right) T(s'|s,a)$$

- Propagation step is a γ-contraction mapping
  - Procedure converges to V^*

Policy Iteration

- But we can have an intermediate policy:
  - Start with some arbitrary $Q_0(\cdot, \cdot)$
  - Loop the following:
    - Compute a greedy policy w.r.t. Q_t:
      $$\pi(s, a) = \arg\max_a Q_t(s, a)$$
    - Compute policy value V^π
    - Compute
      $$Q_{t+1}(s, a) = \sum_{s'} \left(R(s', a, s) + \gamma V^\pi(s')\right) T(s'|s, a)$$
- Converges being a contraction mapping as well

Markov games

- State may be subject to effects by more than one agent
- Multiagent Markovian Environment $< S, s_0, \{A_i\}_{i=1}^N, T >$
  - S and $s_0 \in S$ are the state space and initial state
  - A_i is the space of i'th agent actions
  - $T : S \times A \times S \rightarrow [0, 1]$, where $A = \bigotimes A_i$.
    $T(s', a, s)$ is the probability that state will change from s to s' if joint action $a = (a_1, ..., a_N)$ is taken
- Markov Game is then $< S, s_0, \{A_i\}_{i=1}^N, T, \{R_i\}_{i=1}^N >$
  - $R_i : S \times A \rightarrow \mathbf{R}$, where $A = \bigotimes A_i$
  - Usually discount accumulated

Policy profile

- For regular games we had a mixed strategy profile $\pi = (\pi_1, ..., \pi_N)$
  - $\pi(a) = \prod \pi_i(a_i)$
- For Markov games we define a joint policy profile $\pi = (\pi_1, ..., \pi_N)$
  - $\pi(s, a) = \prod \pi_i(s, a_i)$
- Notice that a policy of an individual agent may be "pure"
  - For each $s \in S$ exists a single $a_i \in A_i$ so that $\pi(s, a_i) = 1$

Minimax solution

- For $N = 2$ and $R_1 = -R_2$ we can formulate a minimax solution
  - Let $V(s)$ be expected reward for the optimal policy starting at state $s \in S$
  - Let $Q(s, a_1, a_2)$ the expected reward for the optimal policy if at first agents perform (a_1, a_2)
- Then system of equations holds:
  - $V(s) = \max_{\pi} \min_{a_2} \sum_{a_1 \in A_1} Q(s, a_1, a_2)\pi(a_1)$
  - $Q(s, a_1, a_2) = R(s, a_1, a_2) + \gamma \sum_{s' \in S} T(s', a_1, a_2, s)V(s')$

Equilibrium solution

- Given the estimate of quality $Q(s,a)$ one can define equilibrium
- Policy profile $\pi = (\pi_1, ..., \pi_N)$ is an equilibrium if for any $\pi' = (\pi'_i, \pi_{-i})$

$$\sum_{a \in A} \pi(s,a) Q_i(s,a) \geq \sum_{a \in A} \pi'(s,a) Q_i(s,a)$$

Background: POMDP

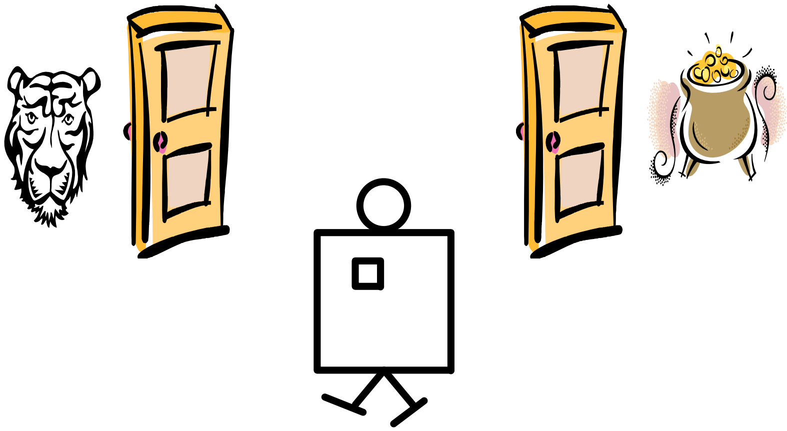
Decision-making in single agent complex domains:
Partially **O**bservable **M**arkov **D**ecision **P**rocess

Single agent Tiger problem (digression from search & rescue)

Task: Maximize collection of gold over a finite or infinite number of steps while avoiding tiger

Tiger emits a growl periodically (GL or GR)

Agent may listen or open doors (L, OL, or OR)



Partially observable environment

- A partially observable Markovian environment $< S, s_0, A, T, \Omega, O >$
  - S state space of the world, s_0 is the initial state
  - A is a set of actions available to the agent
  - $T : S \times A \times S \rightarrow [0, 1]$ is the transition function
  - Ω is the set of all possible observations
  - $O : \Omega \times S \times A \times S \rightarrow [0, 1]$ is the observability function.
    - $O(o|s', a, s)$ is the probability that the agent will observe o if it performed a and the world shifted from s to s'.

Background: POMDP

- Question 1: How rich should S be?
  Answer: As much as you can

- Question 2: What if other agents are present?

- Problem

"... there is currently no good way to combine game theoretic and POMDP control strategies."

- Russell and Norvig
AI: A Modern Approach, 2nd Ed.

Background: POMDP

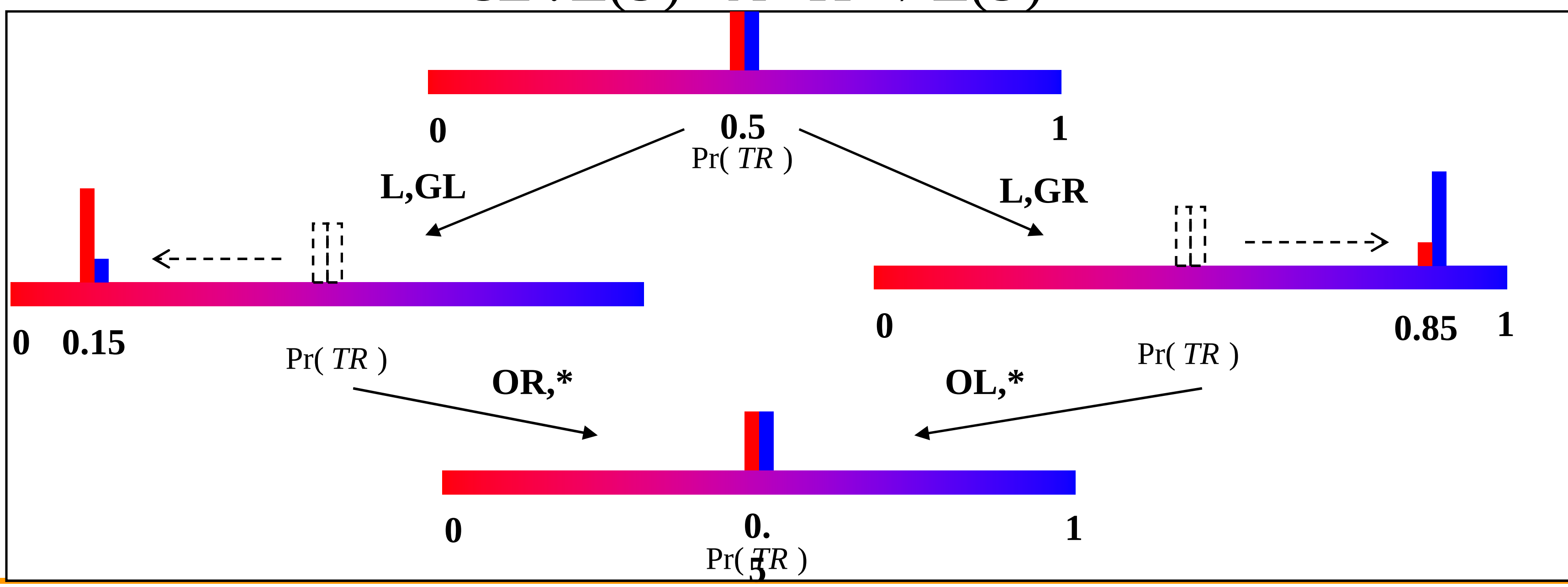
Steps to compute a strategy (policy)

1. Model of the decision making situation:

$$\left\langle S, A_i, \Omega_i, O_i, T_i, R_i, OC_i \right\rangle$$

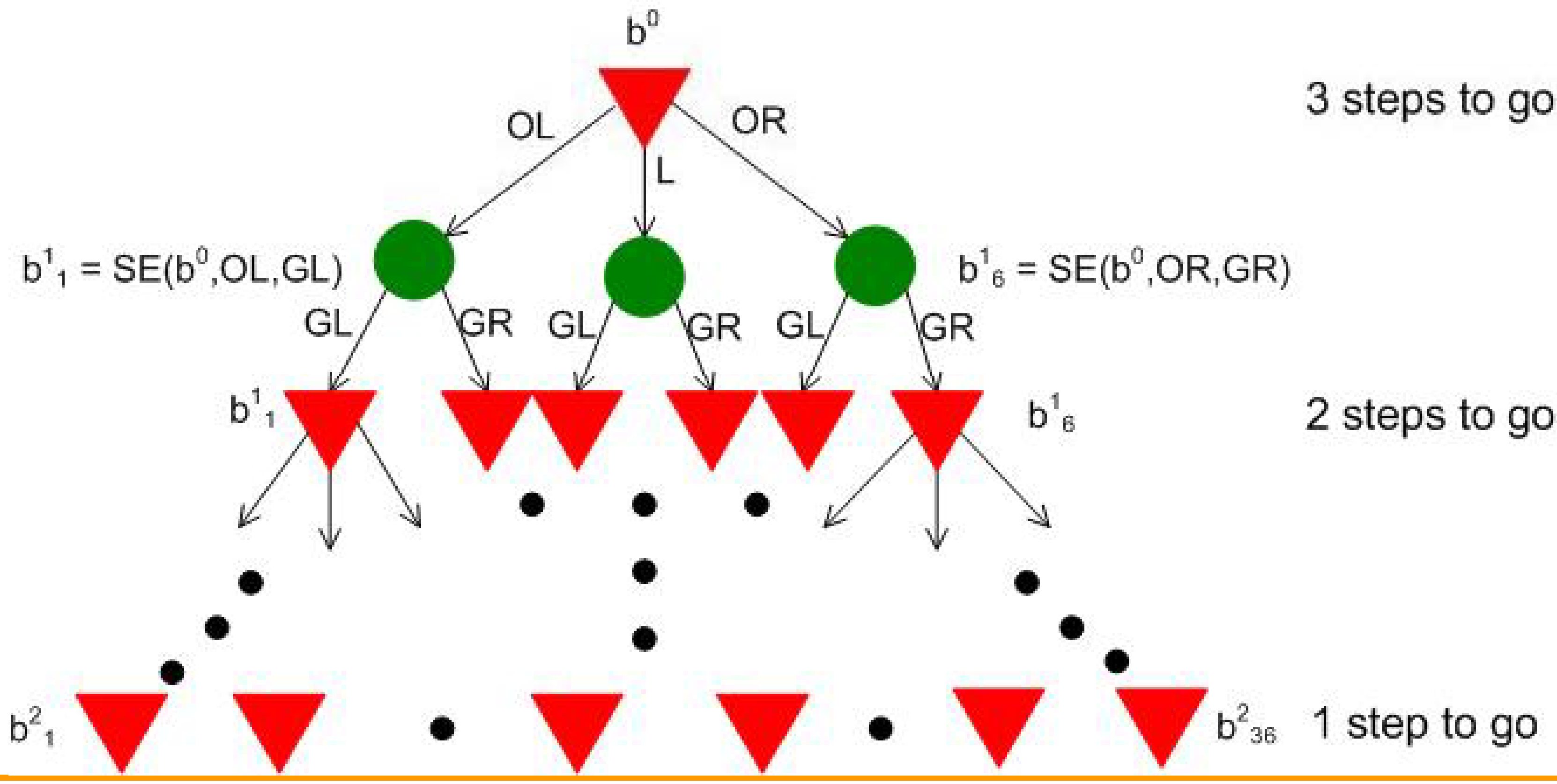
2. Update beliefs:

$$SE : \Delta(S) \times A \times \Omega \rightarrow \Delta(S)$$



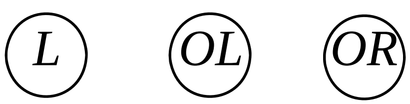
Background: POMDP

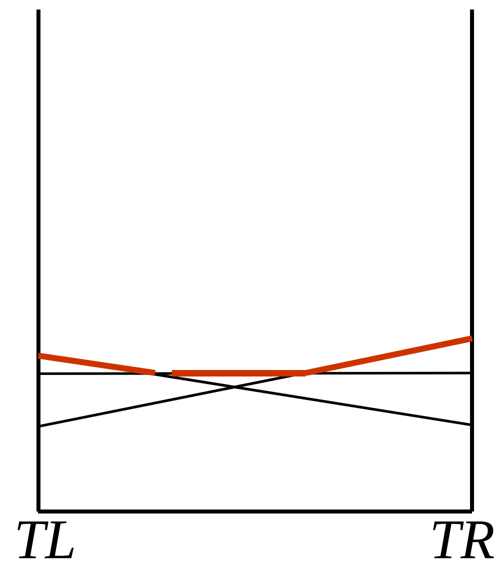
3. Optimal policy computation:
   - Build the look ahead reachability tree
   - Dynamic programming (DP)



Background: POMDP

Dynamic Programming in POMDPs

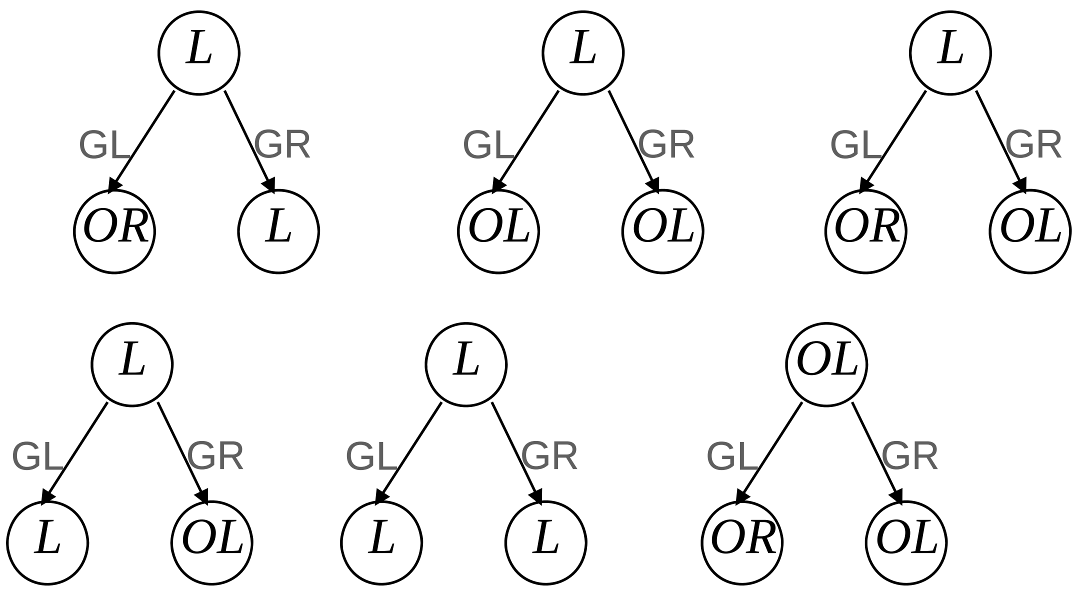


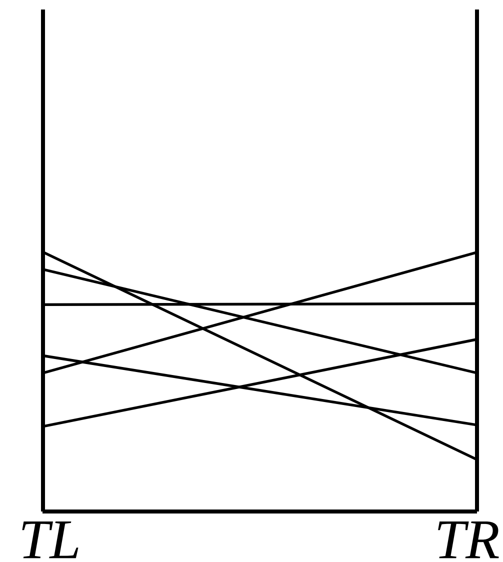


Background: POMDP

DP in POMDPs

Number of policy trees is exponential in observations and doubly exponential in horizons!

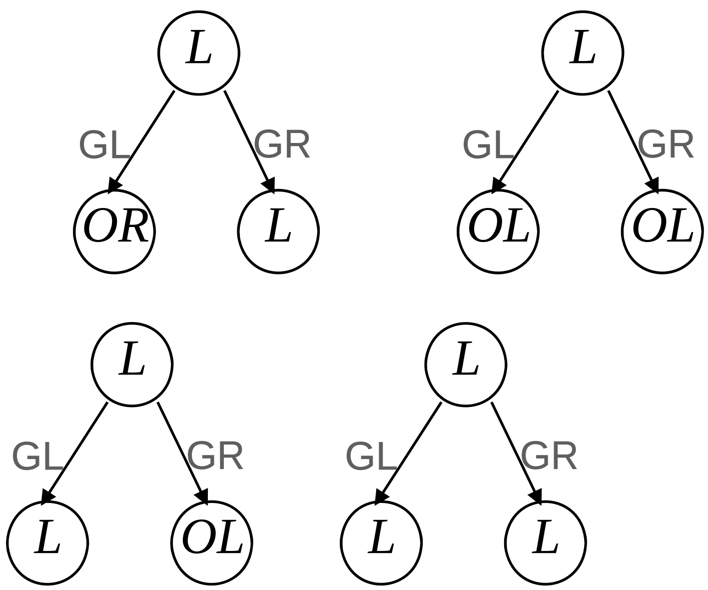


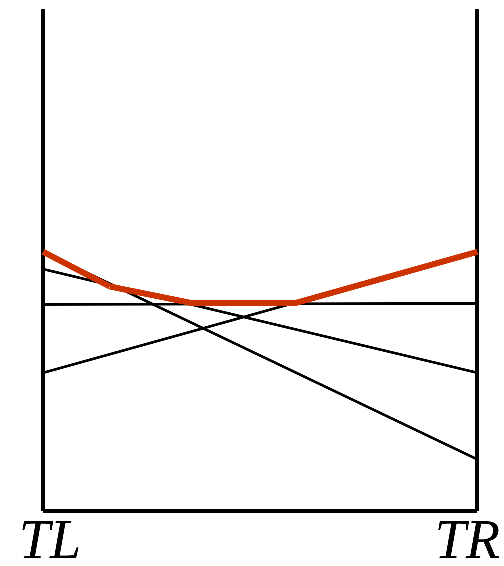


Background: POMDP

DP in POMDPs

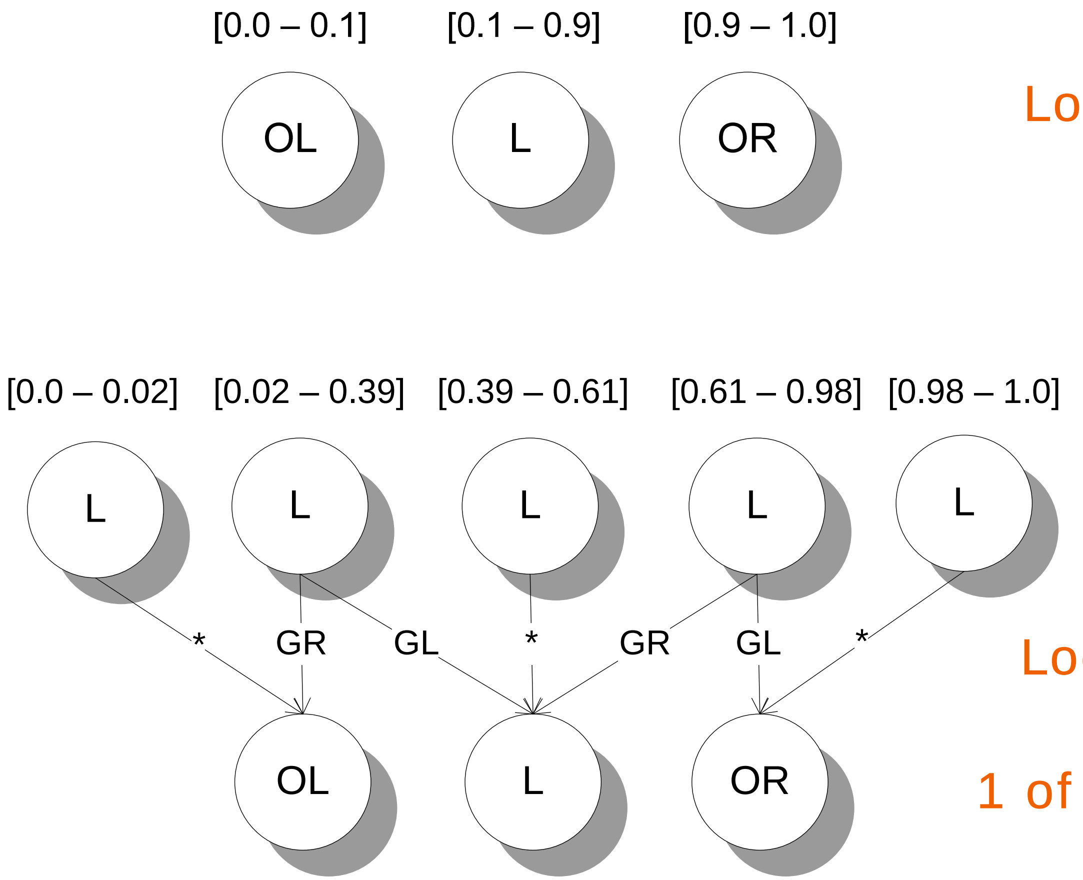
Prune suboptimal policy trees





Background: POMDP

Policies in the tiger problem



Look ahead 1 step (horizon 1)

Look ahead 2 steps (horizon 2)
1 of 4 different policies

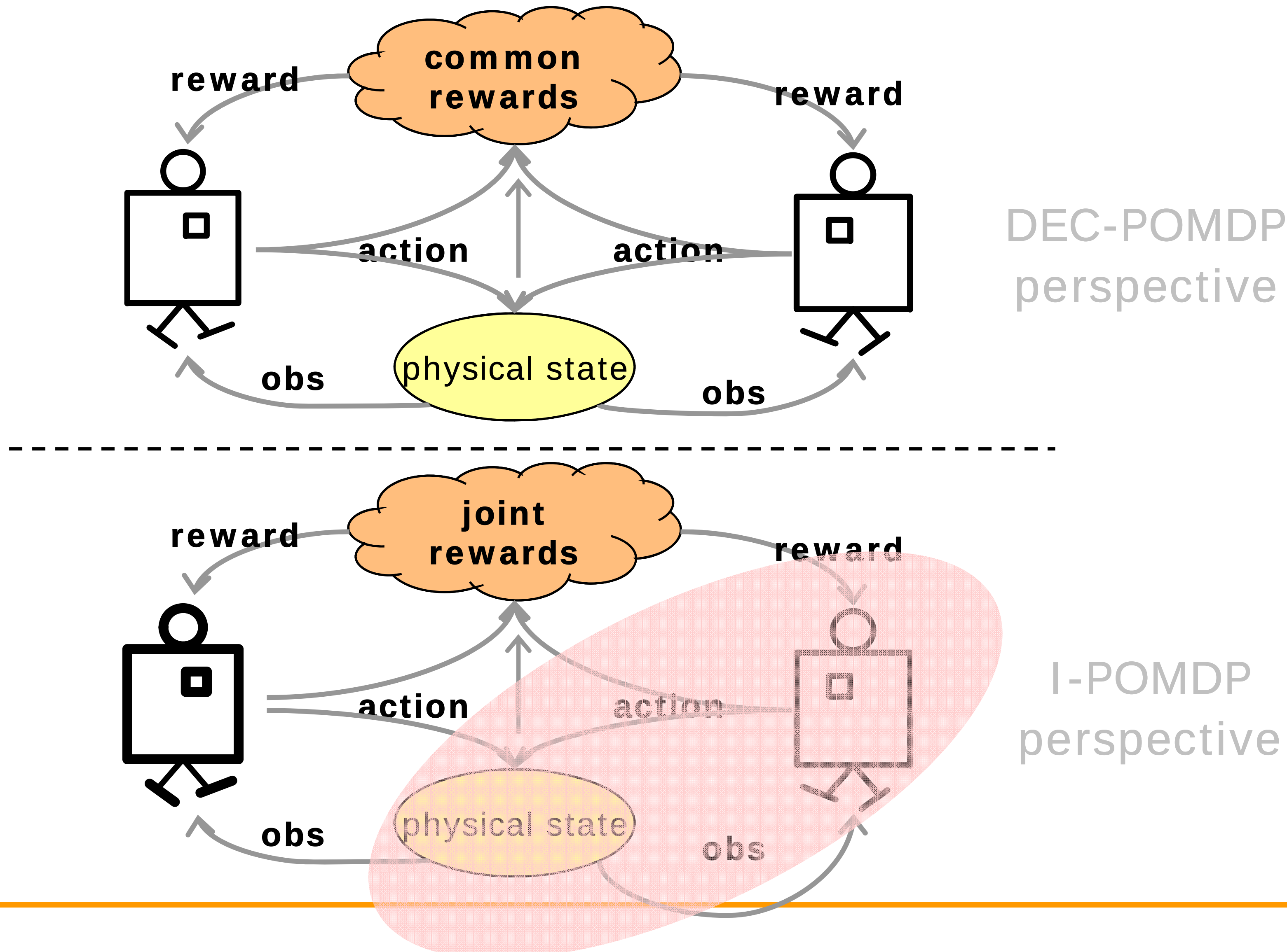
Partially Observable SGs

- Environment $< S, s_0, \{A_i\}_{i=1}^N, \{\Omega_i\}_{i=1}^N, O, T >$
  - S and $s_0 \in \Delta(S)$ are the state space and initial state
  - A_i is the space of i'th agent actions
  - $T : S \times A \times S \rightarrow [0, 1]$ is the state transition function
  - Ω_i is the i'th agent observations space
  - $O : \Omega \times S \times A \times S \rightarrow [0, 1]$ is the observability function, where $\Omega = \bigotimes \Omega_i$
- A POSG is then $< S, s_0, \{A_i\}_{i=1}^N, \{\Omega_i\}_{i=1}^N, O, T, \{R_i\}_{i=1}^N >$
  - $R_i : S \times A \rightarrow \mathbf{R}$, where $A = \bigotimes A_i$
  - Usually discount accumulated

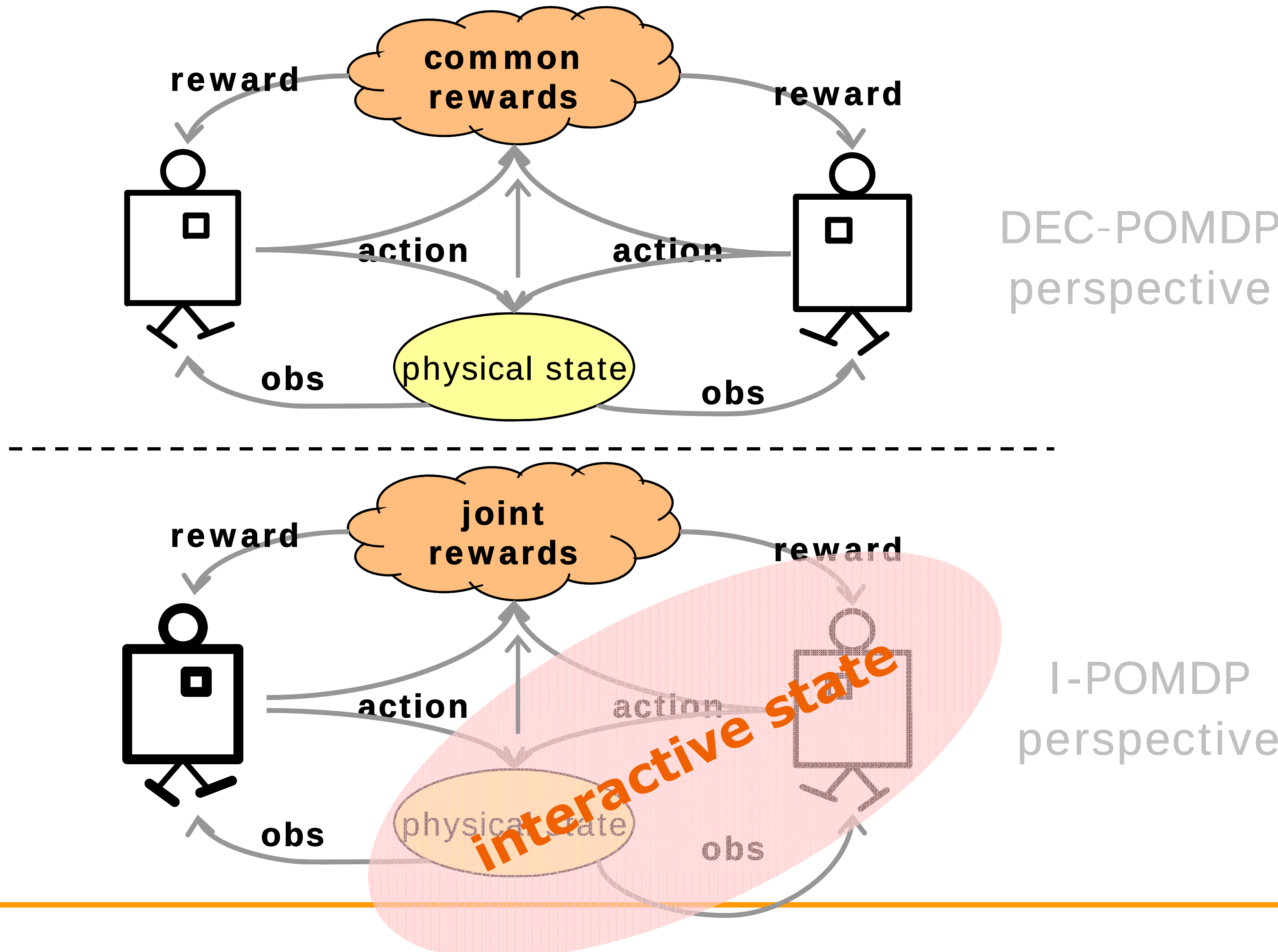
Roadmap: which POSG, if any?

- Classification
  - Based on reward properties
    - E.g. if $\forall i, j \;\; R_i = R$, it is a team game: DEC-POMDP
  - Based individual observability
  - Based on state space structure
- POSGs are **not** all encompasing
  - E.g. reward is internal to agents

DEC-POMDP and I-POMDP



DEC-POMDP and I-POMDP



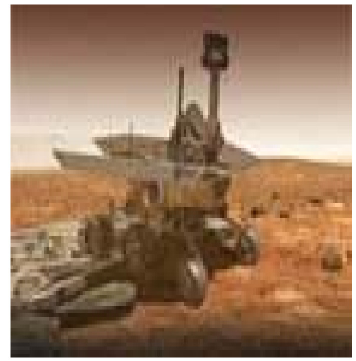
I-POMDP

Key ideas

- Include possible behavioral models of other agents in the state space. Agent's beliefs are distributions over the physical state and models of others
  - Intentional (types) and subintentional models
- Intentional models contain beliefs. Beliefs over models give rise to interactive belief systems
  - Interactive epistemology, recursive modeling
- Finitely nested belief system as a computable approximation of the interactive belief system
- Compute best response to agent's belief (subjective rationality)

Applications

- Robotics
  - Planetary exploration
    - Surface mapping by rovers
    - Coordinate to explore pre-defined region optimally
    - Uncertainty due to sensors
  - Robot soccer
    - Coordinate with teammates and deceive opponents
    - Anticipate and track others' actions



Spirit

Opportunity

RoboCup Competition

I-POMDP

Definition of a finitely nested I-POMDP of strategy level l for agent *i* in a 2 agent setting

$$\langle IS_{i,l}, A, T_i, \Omega_i, O_i, R_i, OC_i \rangle$$

$IS_{i,l}$ is the set of interactive states

$$IS_{i,l} = S \times M_{j,l-1} \quad where \quad M_{j,l-1} = \Theta_{j,l-1} \cup SM_j$$

$\theta_{j,l-1} = \langle b_{j,l-1}, A, T_j, \Omega_j, O_j, R_j, OC_j \rangle$ and Bayes rational

I-POMDP

Definition of a finitely nested I-POMDP of strategy level l for agent *i* in a 2 agent setting

$$\left\langle IS_{i,l}, A, T_i, \Omega_i, O_i, R_i, OC_i \right\rangle$$

$IS_{i,l}$ is the set of interactive states

A is the set of joint actions

T_i is the transition function defined on the physical state (beliefs of others cannot be directly manipulated)

Ω_i is the set of observations of agent *i*

O_i is the observation function (beliefs of others are not directly observable)

R_i is the reward function of agent *i*

Interactive beliefs in I-POMDP

- *"In interactive contexts […], it is important to take into account not only what the players believe about substantive matters […] but also what they believe about the beliefs of other players."*
- *"One specifies what each player believes about the substantive matters, about the beliefs of others about these matters, about the beliefs of others about the beliefs of others, and so on ad infinitum."*

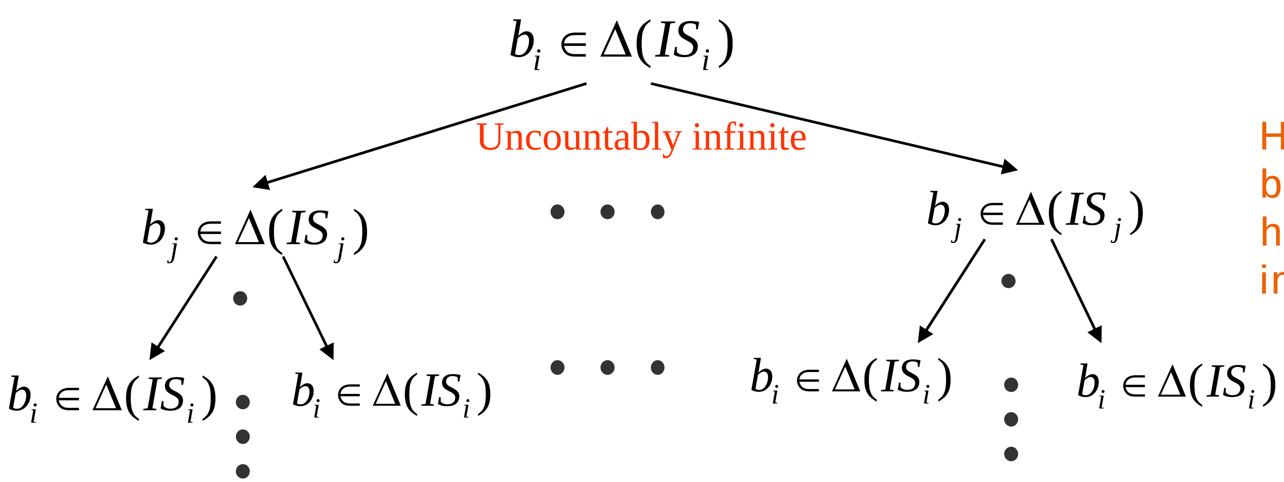
  \- Robert J. Aumann

- New concept: Interactive beliefs
- New approach to game theory: Epistemic, decision analytic

Interactive beliefs in I-POMDP

Agent *i*'s belief is a distribution over the physical state and models of j

$$
\begin{aligned}
b_i \in \Delta(IS_i) &= \Delta(S \times M_j) \\
&= \Delta(S \times \{\langle B_j \times \widehat{\Theta}_j \rangle \cup SM_j\})
\end{aligned}
$$



Hierarchical belief systems have been explored in game theory

Observation

- Amount of information in interactive belief hierarchy is finite
  - Information content decreases asymptotically with the number of levels

- Question 1: How many levels should we include?

  Answer: As many as we can

- Can one work with infinite levels?

  Answer: Yes, in some special cases

Observation

- Minimax in Chess game
  - Model of agent’s possible moves
  - Model the other player’s possible responses
    - Assume she is rational (is she?)
  - Model the other player modeling the agent’s possible responses
    - Assume she believes agent is rational (does she?)
  - Model further …
    - Assume that she believes that agent believes that she is rational …
- Include as much detail and levels as you can

I-POMDP

- Integrate models of others in a decision-theoretic framework
  - An important model is a POMDP describing an agent – it includes all factors relevant to agent's decision making. These are **intentional models (BDI)**
  - Represent uncertainty by maintaining beliefs over the state and models of other agents. This gives rise to interactive belief systems
    - interactive epistemology
  - When no other agents are present beliefs become "flat" and classical POMDP results
  - Computable approximation of the interactive beliefs: finitely nested belief systems
    - infinitely nested beliefs are computable if there is common knowledge – Nash equilibria

Belief update in I-POMDP

- Formalization

$$Pr(is^t|a_i^{t-1}, b_{i,l}^{t-1}) = \beta \sum_{IS^{t-1}:\widehat{m}_j^{t-1}=\widehat{\theta}_j^t} b_{i,l}^{t-1}(is^{t-1})$$
$$\times \sum_{a_j^{t-1}} Pr(a_j^{t-1}|\theta_{j,l-1}^{t-1}) O_i(s^t, a_i^{t-1}, a_j^{t-1}, o_i^t)$$
$$\times T_i(s^{t-1}, a_i^{t-1}, a_j^{t-1}, s^t) \sum_{o_j^t} O_j(s^t, a_i^{t-1}, a_j^{t-1}, o_j^t)$$
$$\times \tau(SE_{\widehat{\theta}_j^t}(b_{j,l-1}^{t-1}, a_j^{t-1}, o_j^t) - b_{j,l-1}^t)$$

Belief update in I-POMDP

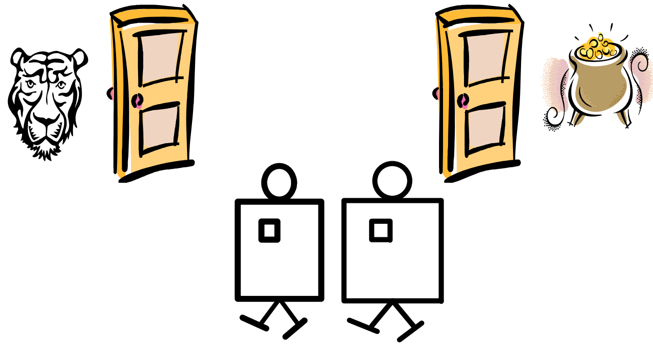
Multiagent Tiger problem

Task: Maximize collection of gold over a finite or infinite number of steps while avoiding tiger

Each agent hears growls as well as creaks (S, CL, or CR)

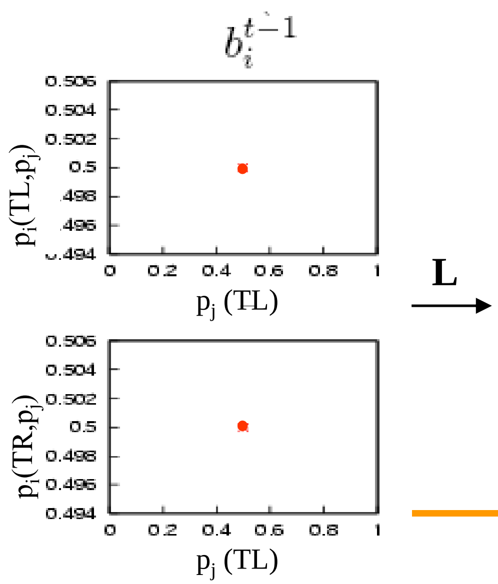
Each agent may open doors or listen

Each agent is unable to perceive other’s observation



agents *i* & *j*

Understanding the I-POMDP (level 1) belief update



Belief update in I-POMDP

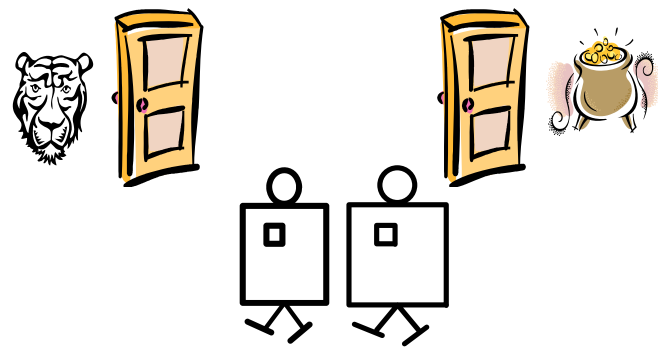
Multiagent Tiger problem

Task: Maximize collection of gold over a finite or infinite number of steps while avoiding tiger

Each agent hears growls as well as creaks (S, CL, or CR)

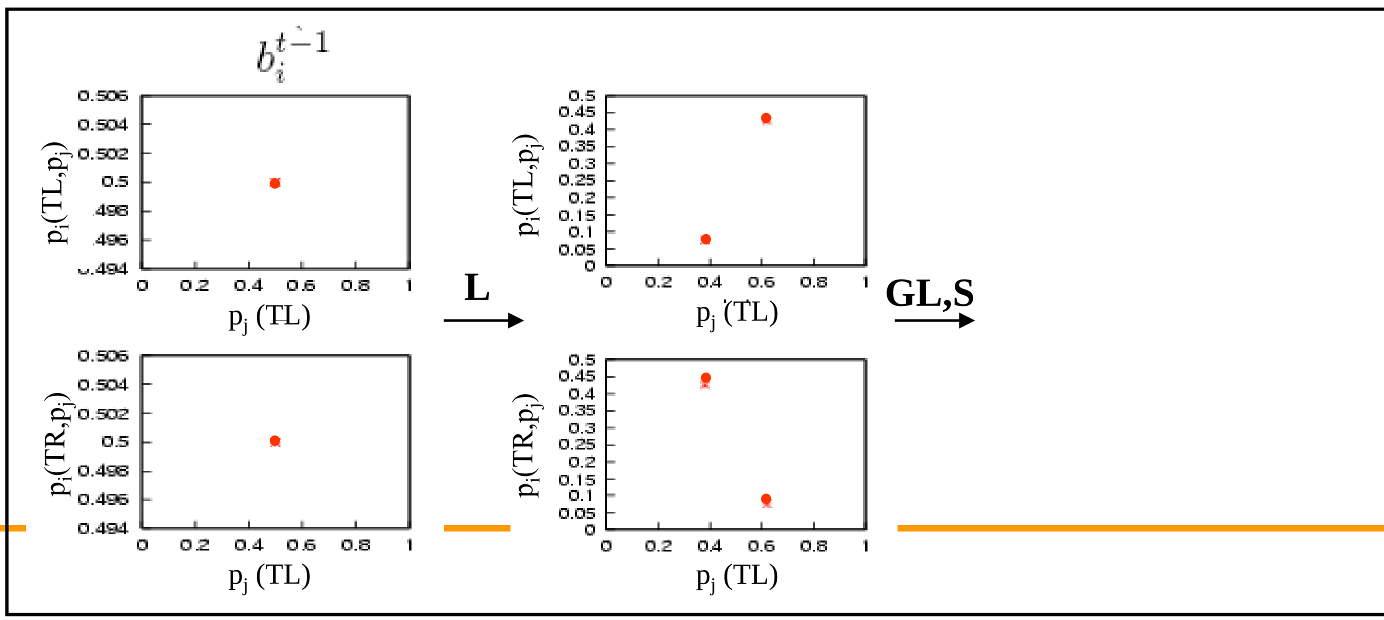
Each agent may open doors or listen

Each agent is unable to perceive other's observation



agents i & j

Understanding the I-POMDP (level 1) belief update



Belief update in I-POMDP

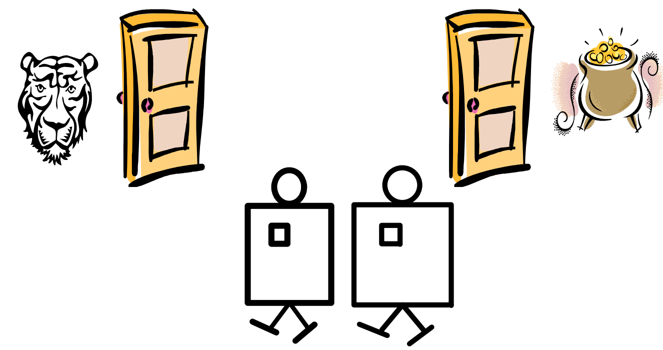
Multiagent Tiger problem

Task: Maximize collection of gold over a finite or infinite number of steps while avoiding tiger

Each agent hears growls as well as creaks (S, CL, or CR)

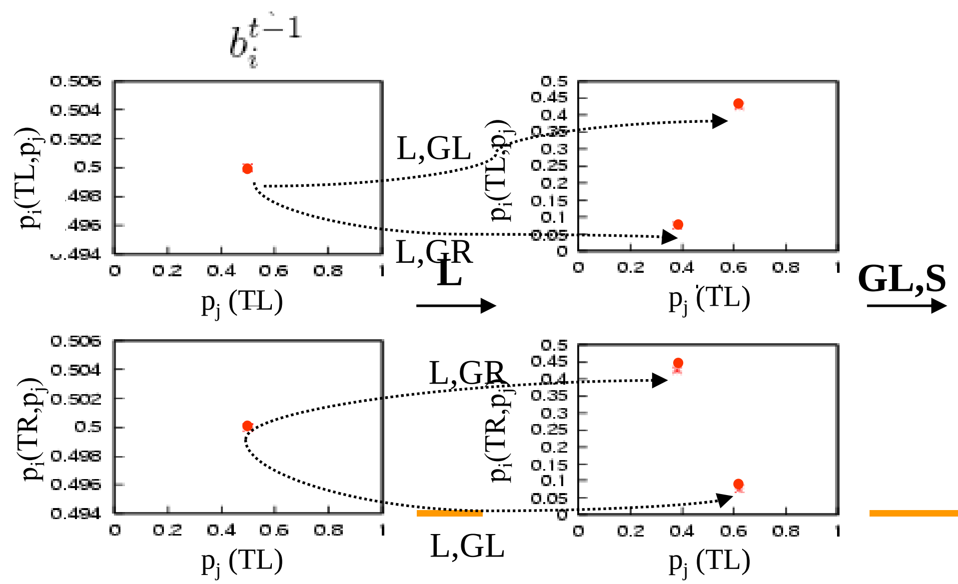
Each agent may open doors or listen

Each agent is unable to perceive other's observation



agents *i* & *j*

Understanding the I-POMDP (level 1) belief update



Belief update in I-POMDP

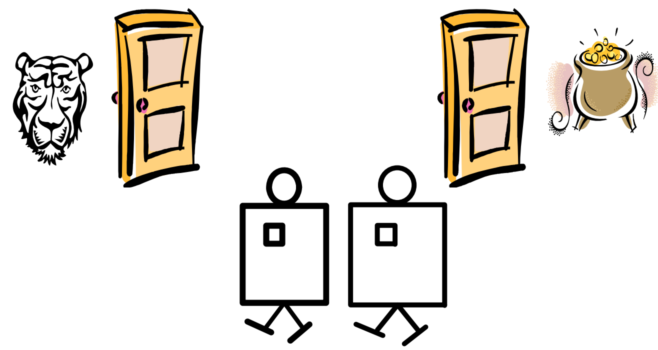
Multiagent Tiger problem

Task: Maximize collection of gold over a finite or infinite number of steps while avoiding tiger

Each agent hears growls as well as creaks (S, CL, or CR)

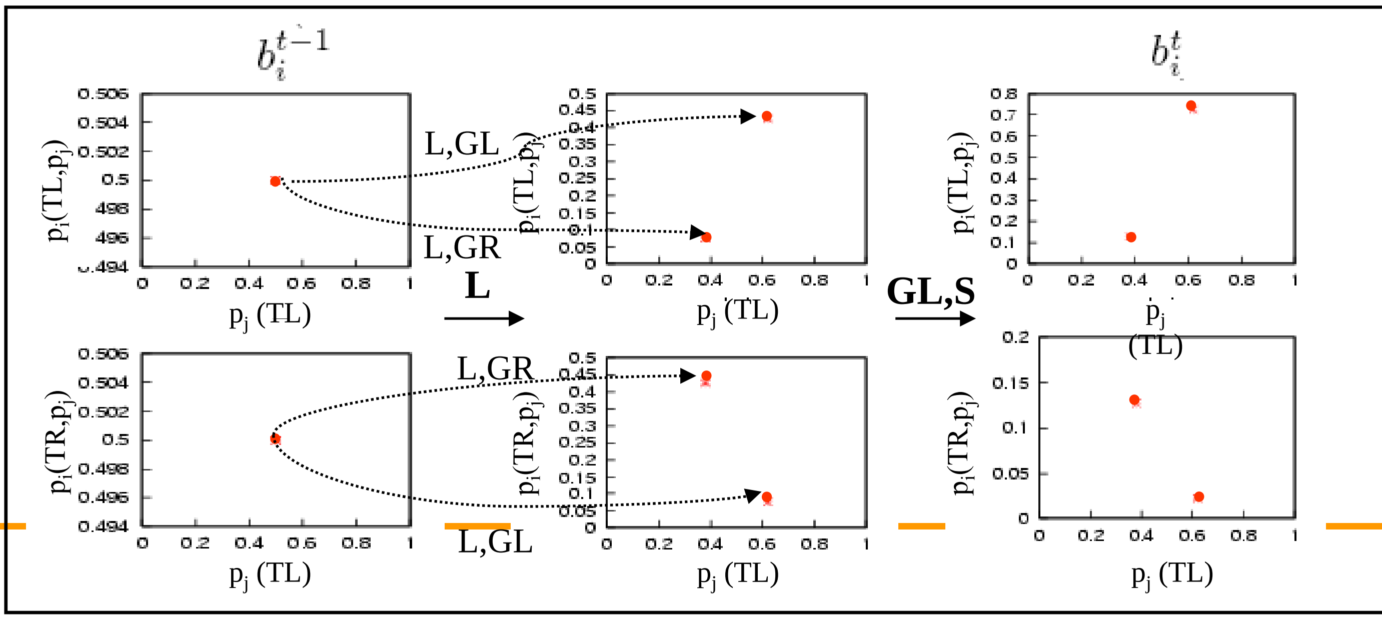
Each agent may open doors or listen

Each agent is unable to perceive other's observation



agents *i* & *j*

Understanding the I-POMDP (level 1) belief update



DP in I-POMDP

Recurse through levels beginning with level 0

Agent j
level 0 models of horizon 1
(assumes agent *i* is noise)

a_2 a_1 a_1 a_1 a_2 a_1 a_2

DP in I-POMDP

Best response to level 1 belief at horizon 1

Agent i
level 1

a_1

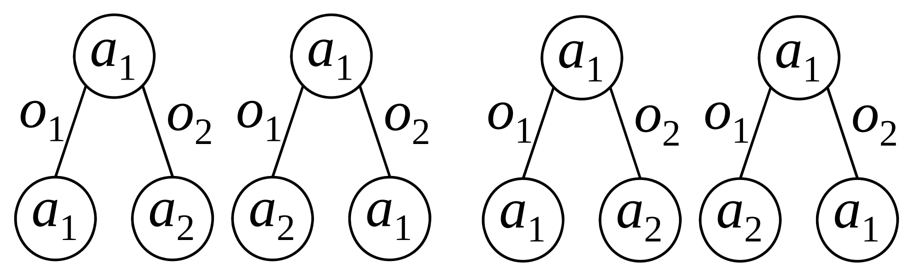
Agent j
level 0 models of horizon 1

a_2 a_1 a_1 a_1 a_2 a_1 a_2

DP in I-POMDP

Agent i
level 1

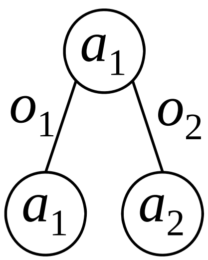
Agent j
level 0 models of horizon 2



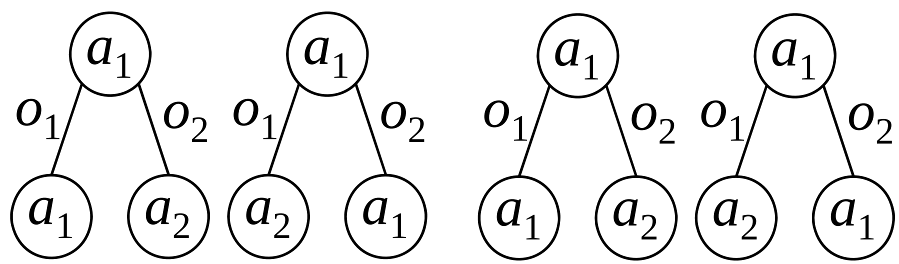
DP in I-POMDP

Best response to level 1 belief at horizon 2

Agent i
level 1



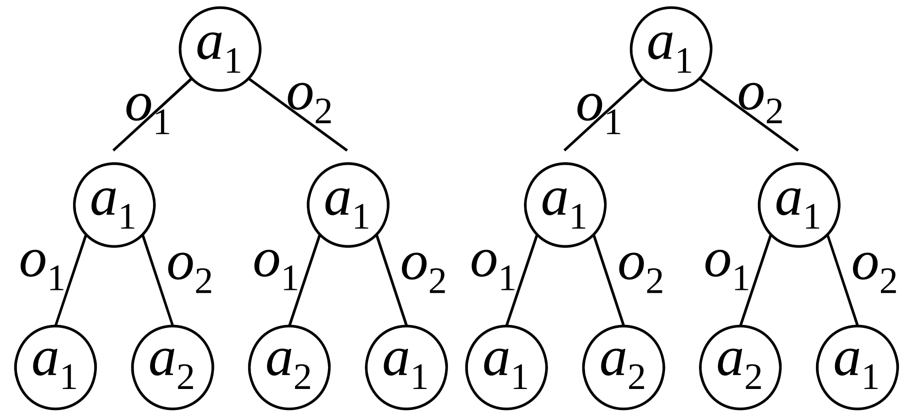
Agent j
level 0 models of horizon 2



DP in I-POMDP

Agent i
level 1

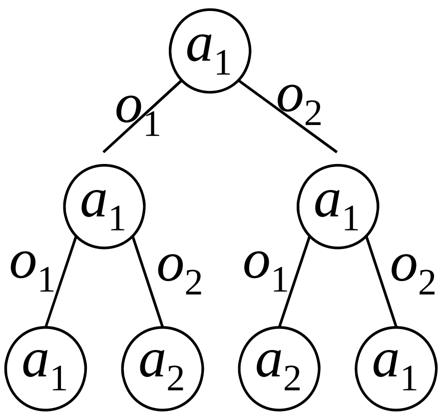
Agent j
level 0 models of horizon 3



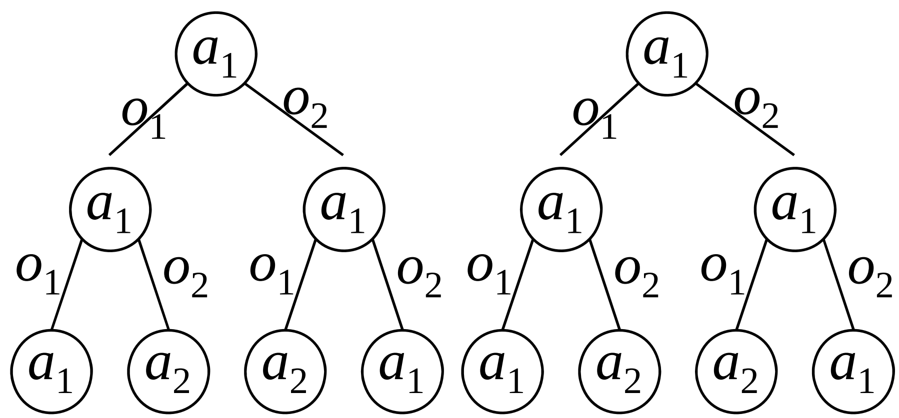
DP in I-POMDP

Best response to level 1 belief at horizon 3

Agent i
level 1



Agent j
level 0 models of horizon 3



POMDPs and I-POMDPs

- Beliefs – probability distributions over states are sufficient statistics
  - They fully summarize the information contained in any sequence of observations
- Solving POMDPs is hard (P-space)
  - We need approximations (e.g., particle filtering)
- Solving I-POMDPs is at least as hard
  - An approximation: interactive particle filtering
- If recursion does not terminate, look for fixed points

Summary of I-POMDPs

- **I-POMDPs:** A framework for decision making in uncertain multiagent settings
- Analogous to POMDPs but with an enriched state space
  - interactive beliefs
- Uses decision-theoretic solution concept
  - MEU
- For infinitely nested beliefs, look for fixed points
- Intractability of I-POMDPs
  - Curse of dimensionality: belief space complexity
  - Curse of history: policy space complexity
- **Approximation 1: Interactive Particle Filter**
  - Randomized algorithm for approximating the nested belief update
  - Partial error bounds
- **Approximation 2: Interactive Influence Diagrams**

Human-Agent Collaboration

- Possible to create a training tool for human emergency response teams.
  - E.g. firefighter managers have been trained using RoboCup Rescue.
- Emergency protocols allow a stochastic model of humans interacting with a simulated environment.
  - Can it be used to devise a flexible training environment?
  - How can we diversify the experience to provide a sufficient span of scenarios?
  - Can a certain degree of surprise be ensured?

Interactive simulations

- Interaction is a sequence of complex events which are
  - extended in time
  - have a component hidden from the human player
- Surprise can be achieved by
  - Exposition of information contrary to the known
    - Find that the building is not abandoned
  - Sequencing of events that require polar response
    - False report of a fire in the North followed by a report that it is in the South

Interactive simulations

- Interaction is a sequence of complex events which are
  - extended in time
  - have a component hidden from the human player
- Surprise can be achieved by
  - Exposition of information contrary to the known
    - Find that the building is not abandoned
  - Sequencing of events that require polar response
    - False report of a fire in the North followed by a report that it is in the South
- How do we produce these sequences?

Interactive simulations

- Interaction is a sequence of complex events which are
  - extended in time
  - have a component hidden from the human player
- Surprise can be achieved by
  - Exposition of information contrary to the known
    - Find that the building is not abandoned
  - Sequencing of events that require polar response
    - False report of a fire in the North followed by a report that it is in the South
- How do we produce different sequences?

Interactive simulations

- Interaction is a sequence of complex events which are
  - extended in time
  - have a component hidden from the human player
- Surprise can be achieved by
  - Exposition of information contrary to the known
    - Find that the building is not abandoned
  - Sequencing of events that require polar response
    - False report of a fire in the North followed by a report that it is in the South
- How do we produce different sequences?
  - Interactive simulations $\equiv$ dynamic narratives

Trajectory Distribution: Intuition

- Markovian environment representation $< S, A, T >$
  - States are plot points experienced by a player
  - Actions are effects external to the player
  - State transitions are plot connections

Trajectory Distribution: Intuition

- Markovian environment representation $< S, A, T >$
  - States are plot points experienced by a player
    - A firefighter discovers a new fire hazard
    - Police finds a new witness
  - Actions are effects external to the player
  - State transitions are plot connections

Trajectory Distribution: Intuition

- Markovian environment representation $< S, A, T >$
  - States are plot points experienced by a player
  - Actions are effects external to the player
  - State transitions are plot connections

Trajectory Distribution: Intuition

- Markovian environment representation $< S, A, T >$
  - States are plot points experienced by a player
  - Actions are effects external to the player
    - A witness approaches the firefighter
    - A bank robbery occurs
  - State transitions are plot connections

Trajectory Distribution: Intuition

- Markovian environment representation $< S, A, T >$
  - States are plot points experienced by a player
  - Actions are effects external to the player
  - State transitions are plot connections

Trajectory Distribution: Intuition

- Markovian environment representation $< S, A, T >$
  - States are plot points experienced by a player
  - Actions are effects external to the player
  - State transitions are plot connections
    - Subject to the player’s behaviour (stochasticity)
    - Subject to the narrator’s decisions (actions)

Trajectory Distribution: Intuition

- Markovian environment representation $< S, A, T >$
  - States are plot points experienced by a player
  - Actions are effects external to the player
  - State transitions are plot connections

Trajectory Distribution: Intuition

- Markovian environment representation $< S, A, T >$
  - States are plot points experienced by a player
  - Actions are effects external to the player
  - State transitions are plot connections
- A story is a trajectory over plot points (states)
- Trajectory distribution means that a different story is told every time

Example – Fire Chief game

- A Fire Chief manages 3 firefighter teams
- Consider three stories:
  - Story 1
  - “Yesterday a firefighter Team A has been withdrawn from the Toy Factory fire and sent to the Docks. As your correspondent has later discovered, the Docks housed dangerous materials, which led to the infamous explosion and the subsequent perish of Team A.”

Example – Fire Chief game

- A Fire Chief manages 3 firefighter teams
- Consider three stories:
  - Story 2
  - “Earlier today, following an anonymous tip, the Fire Chief sent both Team A and Team B to the Docks, leaving only Team C to handle the fire in our beloved Toy Factory. However, this controversial decision proved to be prudent, since it has prevented the explosion of dangerous chemicals in the Docks.”

Example – Fire Chief game

- A Fire Chief manages 3 firefighter teams
- Consider three stories:
  - Story 3
  - “Our ancient Toy Factory sustained yesterday irrecoverable damage due to the fire that spread from its storage rooms. All three of our firefighter teams where at the time deployed at the Docks, where a minor chemicals leak was handled by one of them. As a result, by the time they arrived at the Toy Factory the place was engulfed in flames.”

Example – State Space

- Consider the ratios of firefighter teams present to the necessary number of teams: $(x : x^*, y : y^*)$

Example – State Space

- Consider the ratios of firefighter teams present to the necessary number of teams: $(x : x^*, y : y^*)$
- A story is then a trajectory through this state space

Example – State Space

- Consider the ratios of firefighter teams present to the necessary number of teams: $(x : x^*, y : y^*)$
- A story is then a trajectory through this state space
  - Story 1
  - $(3 : 1, 0 : 1)$ – All teams are at the Toy Factory
  - $(2 : 0, 1 : 2)$ – Team A is recalled to the Docks
  - $(2 : 0, 0 : 2)$ – Explosion kills Team A

Example – State Space

- Consider the ratios of firefighter teams present to the necessary number of teams: $(x : x^*, y : y^*)$
- A story is then a trajectory through this state space
  - Story 2
  - $(3 : 2, 0 : 1)$ – All teams are at the Toy Factory
  - $(1 : 1, 2 : 2)$ – Team A and B are set to the Docks
  - $(1 : 0, 2 : 0)$ – Explosion is prevented at the docks

Example – State Space

- Consider the ratios of firefighter teams present to the necessary number of teams: $(x : x^*, y : y^*)$
- A story is then a trajectory through this state space
  - Story 3
  - $(0 : 1, 3 : 1)$ – All teams are at the Docks
  - $(0 : 3, 3 : 0)$ – Docks are safe, Toy Factory ablaze
  - $(3 : 0, 0 : 0)$ – Too late: Toy Factory burned down

Example – Actions and Transitions

- States are the ratios of firefighter teams present to the necessary number of teams: $(x : x^*, y : y^*)$
  - A story is a trajectory through this state space
- Actions are hints and information given to the player
  - Anonymous call about chemicals at the Docks
  - TV coverage of the Toy Factory fire
  - An explosion at the Docks

Example – Actions and Transitions

- States are the ratios of firefighter teams present to the necessary number of teams: $(x : x^*, y : y^*)$
  - A story is a trajectory through this state space
- Actions are hints and information given to the player
  - Anonymous call about chemicals at the Docks
  - TV coverage of the Toy Factory fire
  - An explosion at the Docks
- How do we choose actions to produce Story 1?
  - How do we choose actions so that Story 3 is more likely?

Target Trajectory Distribution MDP

- Given an Markovian environment: $< S, T, A >$ where
  - S is the set of states of the world,
  - A is the set of actions,
  - $T : S \times A \rightarrow \Delta(S)$ is the transition function with $T(s'|s,a)$ being the probability of the world changing from state s to state s' if the action a was applied.
- Can we prefer a specific long term sequence?
  - Can the preference be soft, i.e. a distribution?

TTD-MDP (cont)

- Let $\tau \subset S^+$ be a set of finite sequences of states.
  - We will assume that τ is formed by paths in a tree.
- Let $\mathcal{P}(\cdot)$ be a distribution over τ.
  - $\mathcal{P}$ represents our preferences over various, long-term system developments
- A TTD-MDP is defined by a tuple $<< S, T, A >, \tau, \mathcal{P} >$
  - Notice that a transition function $\mathcal{T} : \tau \times A \rightarrow \Delta(\tau)$ is naturally induced by T.

TTD-MDP: Questions

- Given a TTD-MDP, $< < S, T, A >, \tau, \mathcal{P} >$
- What is the policy $\pi : \tau \rightarrow A$ that induces $\mathcal{P}$?
  - Is it always possible to produce $\mathcal{P}$?
    - No, transition function T may prevent that.

TTD-MDP: Questions

- Given a TTD-MDP, $<<S,T,A>,\tau,\mathcal{P}>$
- What is the policy $\pi : \tau \longrightarrow A$ that induces $\mathcal{P}$?
  - Is it always possible to produce $\mathcal{P}$?
    - No, transition function T may prevent that.
  - How do we measure performance?
    - Information Theory provides a divergence measure between two distributions: Kullback-Leibler divergence

TTD-MDP: Questions

- Given a TTD-MDP, $< < S, T, A >, \tau, \mathcal{P} >$
- What is the policy $\pi : \tau \to A$ that induces $\mathcal{P}$?
  - Is it always possible to produce $\mathcal{P}$?
    - No, transition function T may prevent that.
  - How do we measure performance?
    - Information Theory provides a divergence measure between two distributions: Kullback-Leibler divergence
  - Can the policy be computed on-line?
    - Yes, the structure of τ combined with appropriate performance measure allow that.

TTD-MDP: Further Questions

- Assumes complete observability
  - Active plot point is always known to the narrator
  - Will not hold if the narrator is part of the simulation
- Trajectories are finite
  - What if it's a never-ending story?
  - Can a TTD-like principle be defined for infinite trajectories?
- Single agent
  - What if the simulation includes multiple "narrators"?
  - Can a similar TTD principle be applied for multi-agent simulations?

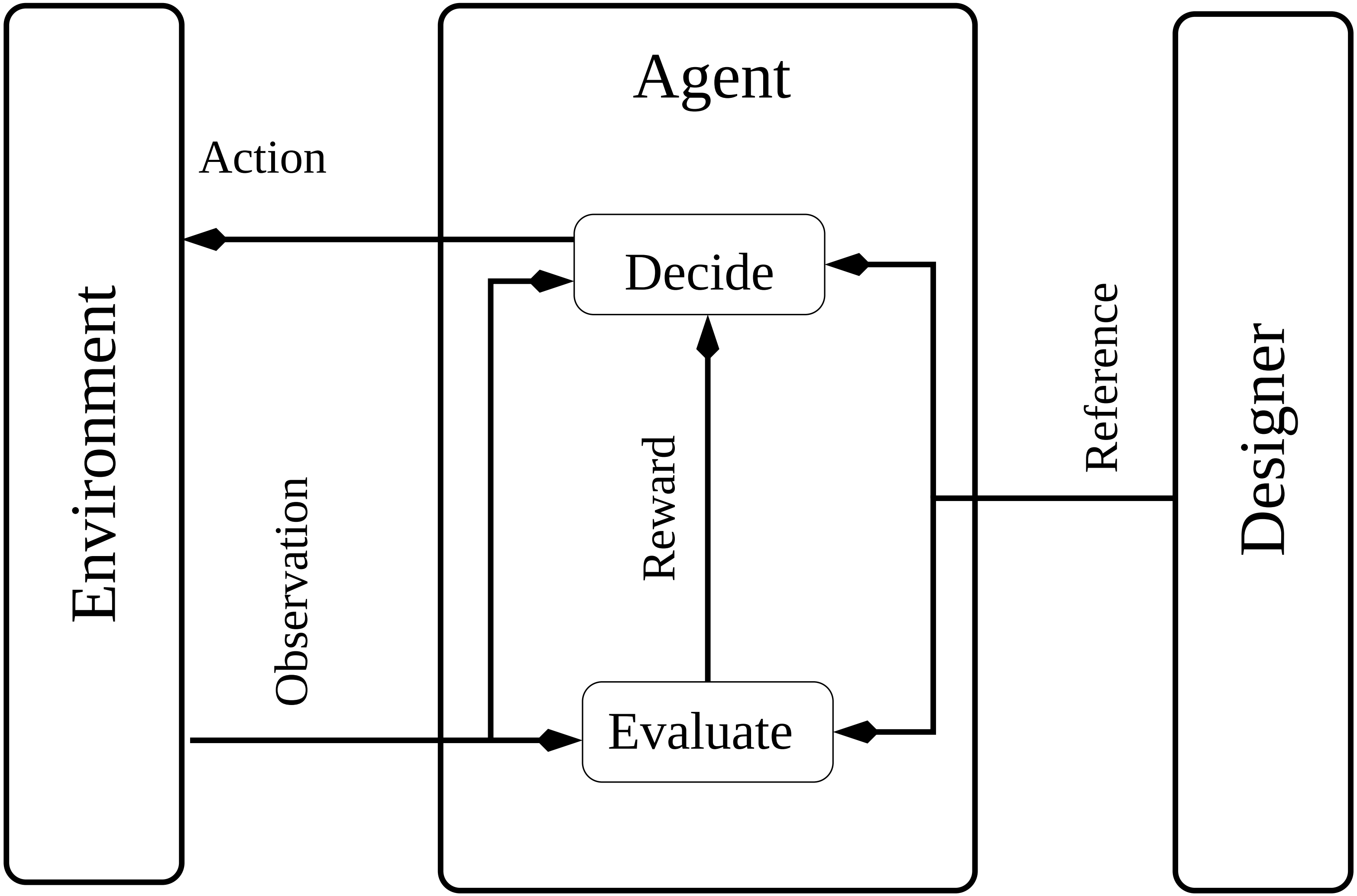
Example

- Two police precincts are fighting organised crime
  - They are unable to catch the leader
    - There are signs of him being in the precinct, but not the exact location
  - They know that increased patrols make him uncomfortable
  - If the leader moves from precinct to precinct, his crime activity is disrupted
- Ideally the police would like to modulate patrols so as to keep the crime leader in constant agitation

Markovian, but not MDP

- Given a partially observable Markovian environment $< S, A, T, O, \Omega >$
- (PO)MDPs define
  - Utility (reward) for state transition as performance estimate
  - Accumulation (averaging) as time extended evaluation
- Is there an alternative?

Egocentric Design



Example (cont)

- Environment $< S, \bigotimes A_i, T, \bigotimes \Omega_i, \{O_i\} >$
  - $S = \{pr_1, pr_2\}$ is the set of precincts
  - $A_i = \{higher, lower\}$ is increasing or decreasing patrols
  - $\Omega_i = S$ is an indicator of leader's presence in the precinct
  - T reflects leader's tendency to move
  - O_i reflects the police capability to gather information
- Reference dynamics is then $\tau(s', s) = \begin{cases} 1 & s \neq s' \\ 0 & otherwise \end{cases}$

Extended Markovian Tracking

- Reference takes the form $\tau^* : S \longrightarrow \Delta(S)$
- Evaluation of the agent
  - Aggregates observations into $\tau^{EMT} : S \longrightarrow \Delta(S)$
  - Internal reward based on the discrepancy between τ^{EMT} and τ^*
- Decision of the agent
  - Predict how actions influence τ^{EMT}
  - Choose actions to minimise the future discrepancy between τ^{EMT} and τ^*

Data Aggregation

- State information $p_t \in \Delta(S)$
  - Given that an agent performed action a and received observation o:
$$p_{t+1}(s) \propto O(o|s,a) \sum_{s'} \mathcal{T}(s|a,s') p_t(s')$$
- Dynamics Estimate $\tau : S \times S \to [0,1]$
  - For τ has to hold $p_{t+1} = p_t * \tau$
  - Make a conservative update:
$$\tau_{t+1} = \arg \min_{\tau : p_{t+1} = p_t * \tau} d(\tau, \tau_t)$$
  - EMT's update is shorthanded $H[p_{t+1} \leftarrow p_t, \tau_t]$

EMT Control

- It is possible to utilise EMT to construct an on-line policy to reproduce a reference dynamics τ^*
- Control loop is composed by
  - Belief update
  - EMT estimation of system development
  - Let $T_a = T(\cdot|a,\cdot)$. Action choice

$$a^* = \arg\min_a D_{KL}(\ H[p_t * T_a \leftarrow p_t, \tau_t] \parallel \ \tau^*)$$

  - Application of a^*.
- But can it be used in a multi-agent setting?

Stigmergy

- Stigmergy is a mechanism of spontaneous, indirect coordination
  - Trace left in the environment by an action stimulates the performance of a subsequent action, by the same or a different agent.
- Assume that two agents choose actions a_1,a_2 and the joint operation (a_1, a_2) is applied on a common system state.
  - In a stigmergic environment observations will provide information on the state dynamics and enable action coordination

Multi-agent EMT

- Given an environment: $< S, \bigotimes A_i, T, \bigotimes O_i, \{\Omega_i\} >$, and a reference dynamics τ^*
- Let each agent run independent EMT based control on the complete actions space $\bigotimes A_i$ as follows:
  - Update beliefs p_t according to T and O_i
  - Compute EMT estimate of system development
  - Compute optimal *joint action*

$$a^* = (a_1^*, ..., a_N^*) = \arg\min_a D_{KL}(\, H[p_t * T_a \leftarrow p_t, \tau_t] \,\|\, \tau^*)$$

  - Apply a_i^*

Stigmergy – example

- Each police precinct will
  - Estimate the apparent crime leader behaviour
  - Predict the effect of a coordinated patrols.
  - Apply the local portion of the joint action

Stigmergy – example

- Each police precinct will
  - Estimate the apparent crime leader behaviour
    - Using crime leader model and EMT
  - Predict the effect of a coordinated patrols.
  - Apply the local portion of the joint action

Stigmergy – example

- Each police precinct will
  - Estimate the apparent crime leader behaviour
  - Predict the effect of a coordinated patrols.
  - Apply the local portion of the joint action

Stigmergy – example

- Each police precinct will
  - Estimate the apparent crime leader behaviour
  - Predict the effect of a coordinated patrols.
    - These joint actions are not necessarily the same
  - Apply the local portion of the joint action

Stigmergy – example

- Each police precinct will
  - Estimate the apparent crime leader behaviour
  - Predict the effect of a coordinated patrols.
  - Apply the local portion of the joint action

Stigmergy – example

- Each police precinct will
  - Estimate the apparent crime leader behaviour
  - Predict the effect of a coordinated patrols.
  - Apply the local portion of the joint action
    - Combined into a joint action different from all player choices

Stigmergy – example

- Each police precinct will
  - Estimate the apparent crime leader behaviour
  - Predict the effect of a coordinated patrols.
  - Apply the local portion of the joint action
- Crime leader responds to the combined joint action leading to stigmergy

Stigmergy – example

- Each police precinct will
  - Estimate the apparent crime leader behaviour
  - Predict the effect of a coordinated patrols.
  - Apply the local portion of the joint action
- Crime leader responds to the combined joint action leading to stigmergy
  - Observations provide a correlation signal
  - Dynamics estimates are correlated
  - Locally computed joint actions will not differ

Stigmergy – example

- Each police precinct will
  - Estimate the apparent crime leader behaviour
  - Predict the effect of a coordinated patrols.
  - Apply the local portion of the joint action
- Crime leader responds to the combined joint action leading to stigmergy
  - Observations provide a correlation signal
  - Dynamics estimates are correlated
  - Locally computed joint actions will not differ
    - too much too frequently
    - in their effect on the dynamics estimate

Stochasticity is Bad

- System is continually changing

Stochasticity is Bad

- System is continually changing
  - No single state trajectory is certain

Stochasticity is Bad

- System is continually changing
  - No single state trajectory is certain
- In partially observable systems

Stochasticity is Bad

- System is continually changing
  - No single state trajectory is certain
- In partially observable systems
  - Can not track a single state trajectory

Stochasticity is Bad

- System is continually changing
  - No single state trajectory is certain
- In partially observable systems
  - Can not track a single state trajectory
  - Concept of system dynamics is needed

Stochasticity is Bad

- System is continually changing
  - No single state trajectory is certain
- In partially observable systems
  - Can not track a single state trajectory
  - Concept of system dynamics is needed
  - Only apparent dynamics can be used

Stochasticity is Good

- System is continually changing
  - No single state trajectory is certain
- In partially observable systems
  - Can not track a single state trajectory
  - Concept of system dynamics is needed
  - Only apparent dynamics can be used

Bibliography

- Game theory
  1. Fundenberg, D. and Tirole, J., *Game theory*. MIT Press (textbook)
  2. Owen, *G., Game theory*. 3rd Edition, Academic Press (textbook)
  3. Binmore, K., *Essays on foundations of game theory.* Pittman, (edited book)
  4. Harsanyi, J. C. (1967). Games with incomplete information played by ‘Bayesian' players. *Management. Science*, 14(3), 159-182 (reference on Bayesian games)
  5. Fudenberg, D., & Levine, D. (1997). *Theory of Learning in Games*. MIT Press (book for fictitious play)

- Regret matching
  1. Hart, S. & Mas-Colell A. (2000). A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5): 1127-1150
  2. Hart, S. (2005). Adaptive heuristics. *Econometrica*, 73(5): 1401-1430

Bibliography

- DEC-MDP
  1. Bernstein, D. S., Givan, R., Immerman, N., & Zilberstein, S. (2002). The complexity of decentralized control of Markov decision processes. *Mathematics of Operations Research*, *27*(4), 819-840 (ref. on complexity of DEC-MDPs)
  2. Becker, R., Zilberstein, S., Lesser, V., & Goldman, C. V. (2003). Transition-independent decentralized markov decision processes. In *Autonomous Agents and Multiagent Systems Conference* (ref. on TI-DEC-MDPs)

- Uncertainty Utilization
  1. Rabinovich, Z. & Rosenschein, J. S. (2005). Multiagent Coordination by Extended Markov Tracking. In *Autonomous Agents and Multiagent Systems Conference* (ref. on EMT)
  2. David L. Roberts, Mark J. Nelson, Charles L. Isbell, Jr., Michael Mateas, and Michael L. Littman. (2006). Targeting Specific Distributions of Trajectories in MDPs. In *Twenty-first Conference on Artificial Intelligence (ref. on TTD-MDPs)*
  3. Sooraj Bhat, David L. Roberts, Mark J. Nelson, Charles L. Isbell, and Michael Mateas. (2007). A Globally Optimal Online Algorithm for TTD-MDPs. In *Autonomous Agents and Multiagent Systems Conference* (ref. on TTD-MDPs)

Bibliography

- DEC-POMDP and specializations

1. Bernstein, D. S., Givan, R., Immerman, N., & Zilberstein, S. (2002). The complexity of decentralized control of Markov decision processes. *Mathematics of Operations Research*, *27*(4), 819-840 (ref. on complexity)
2. Hansen, E., Bernstein, D., & Zilberstein, S. (2004). Dynamic Programming for Partially Observable Stochastic Games. In *Nineteenth National Conference on Artificial Intelligence* (ref. on DP in DEC-POMDPs)
3. Szer, D., & Charpillet, F. (2006). Point-based Dynamic Programming for DEC-POMDPs. In *Twenty-First National Conference on Artificial Intelligence* (ref. on point based DP in DEC-POMDPs)
4. Seuken, S., and Zilberstein, S. (2007) Memory-bounded Dynamic Programming for Decentralized POMDPs. In *International Joint Conference on Artificial Intelligence* (ref. on MBDP in DEC-POMDPs)
5. Nair, R., Tambe, M., Yokoo, M., Pynadath, D., & Marsella, S. (2003). Taming Decentralized Pomdps :Towards Efficient Policy Computation for Multiagent Settings. In *International Joint Conference on Artificial Intelligence* (ref. on MTDP and JESP)
6. Nair R., Varakantham, P., Tambe, M., & Yokoo, M. (2005). Networked Distributed POMDPs: A Synthesis of Distributed Constraint Optimization and POMDPs. In *Twentieth National Conference on Artificial Intelligence* (ref. on ND-POMDPs)

Bibliography

- Interactive POMDP

1. Gmytrasiewicz, P. J., & Doshi, P. (2005). A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research*, *24:49-79* (ref. on I-POMDP)
2. Doshi, P., & Gmytrasiewicz, P. J. (2009). Monte Carlo Sampling Methods for Approximating Interactive POMDPs. *Journal of Artificial Intelligence Research*, *34:297-337* (ref. on PF in I-POMDP)
3. Perez, D., & Doshi, P. (2008). Generalized Point Based Value Iteration for Interactive POMDPs. In Twenty-third Conference on Artificial Intelligence (AAAI) (ref. on PBVI in I-POMDP)
4. Aumann, R. J. (1999). Interactive epistemology i: Knowledge. *International Journal of Game Theory*, *28*, 263-300
5. Brandenburger, A., & Dekel, E. (1993). Hierarchies of beliefs and common knowledge. *Journal of Economic Theory*, *59*, 189-198 (ref. on hierarchical belief systems)